VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2021 Tereza Rumlerová

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

POROVNÁNÍ BAKTERIÁLNÍCH GENOMŮ ZÍSKANÝCH Z DRUHÉ A TŘETÍ GENERACE SEKVENÁTORŮ.

COMPARISON OF BACTERIAL GENOMES OBTAINED FROM SECOND AND THIRD GENERATION SEQUENCING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Tereza Rumlerová**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Markéta Nykrýnová**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Tereza Rumlerová ***ID:*** 211661

Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Porovnání bakteriálních genomů získaných z druhé a třetí generace sekvenátorů.

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma sekvenační technologie, zaměřte se na sekvenování pomocí platforem Illumina Miseq a Oxford Nanopore MinION a následné sestavování genomů. 2) Seznamte se s formátem a strukturou zápisu sekvenačních dat, otestujte kvalitu analyzovaných genomů a sestavte poskytnuté bakteriální genomy získané z FN Brno. 3) Navrhněte metodu pro porovnání bakteriálních genomů získaných z různých sekvenačních dat a dílčí části realizujte. 4) Implementujte navrženou metodu ve vhodném programovacím prostředí. 5) Pomocí navržené metody srovnejte sestavené bakteriální genomy získané z druhé a třetí generace sekvenátorů.

DOPORUČENÁ LITERATURA:

[1] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and Genome Assembly. 2016, 14(5), 265-279. DOI: 10.1016/j.gpb.2016.05.004. ISSN 16720229.

[2] METZKER, Michael L. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010, 11(1), 31-46. DOI: 10.1038/nrg2626. ISSN 1471-0056.

Termín zadání: 8.2.2021 ***Termín odevzdání:*** 28.5.2021

Vedoucí práce: Ing. Markéta Nykrýnová

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Tato bakalářská práce se zabývá sekvenačními technologiemi se zaměřením především na druhou a třetí generaci sekvenátorů a platformy Illumina Miseq a Oxford Nanopore Technologies MinION. Je zde uveden popis sestavování genomů s praktickou ukázkou sestavení 6 genomů bakterie *Klebsiella pneumoniae* poskytnuté FN Brno a testování kvality těchto genomů. Závěrečná část obsahuje teoretický popis a praktickou implementaci vybraných metod pro porovnání sestavených genomů pocházejících z dvou odlišných sekvenačních platforem.

KLÍČOVÁ SLOVA

Sekvenační technologie, sestavování genomu, porovnání bakteriálních genomů

ABSTRACT

This bachelor thesis deals with sequencing technologies with a focus on the second and third generation of sequencers and platforms Illumina Miseq and Oxford Nanopore Technologies MinION. There is a description of an assembly of genomes with an example of the assembly of 6 genomes of a bacteria *Klebsiella pneumoniae* and a quality testing of these genomes. The final part contains a theoretical description and a practical implementation of selected methods for comparison of the assembled genomes, which come from two different sequencing platforms.

KEYWORDS

Sequencing technology, genome assembly, comparison of bacterial genomes

RUMLEROVÁ, Tereza. *Porovnání bakteriálních genomů získaných z druhé a třetí generace sekvenátorů*. Brno, 2021, 86 s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Markéta Nykrýnová

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Porovnání bakteriálních genomů získaných z druhé a třetí generace sekvenátorů“ jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí mé bakalářské práce paní Ing. Markétě Nykrýnové za odborné vedení, vstřícnost a trpělivost při konzultacích a především za podnětné návrhy k práci.

Obsah

Seznam obrázků

Úvod

Tato bakalářská práce je zaměřena na sekvenační technologie, sestavování genomu a metody pro porovnání genomů. Sekvenací se rozumí proces určení primární struktury genetické informace, tedy pořadí nukleotidových bází v sekvencích deoxyribonukleové kyseliny (DNA). K tomuto účelu bylo a stále je vyvíjeno mnoho sekvenačních platforem využívajících různé postupy jak přípravy knihovny, tak samotného čtení jednotlivých bází.

Začátek teoretické části bakalářské práce pojednává právě o těchto technikách a představuje všechny generace sekvenátorů, především však druhou a třetí, které jsou pro tuto práci stěžejní. V druhé kapitole je obsažen proces sestavování genomů ze čtení produkovaných sekvenačními platformami. Jsou zde zmíněny dva, v dnešní době hojně využívané, algoritmy, a to tzv. de Bruijnův graf a metoda využívající překryvu známá jako Over Layout Consensus. Dále následuje popis formátů využívaných při práci se sekvenačními daty a na objasnění struktury jejich zápisu.

Čtvrtá kapitola je počátkem praktické části této bakalářské práce. Nejprve došlo k sestavení kompletních sekvencí genomů bakterie *Klebsiella pneumoniae* získaných z Fakultní nemocnice Brno, kde byly osekvenovány. K dispozici byly dvě sady sekvenačních dat. Jedna pocházela z platformy Illumina Miseq a druhá ze sekvenátoru MinION od společnosti Oxford Nanopore Technologies. Navazující pátá kapitola slouží k otestování kvality poskytnutých genomů. V první řadě je testována kvalita čtení při sekvenaci, dále pak kvalita sestavení kompletních sekvencí.

Závěrečná část bakalářské práce se zabývá metodami pro porovnání bakteriálních genomů. Nejprve je zde obsažen teoretický úvod zaměřený na metody, které je možné využít, a poté následuje jejich praktická implementace. V rámci analýzy sestavených sekvencí bylo provedeno zarovnání stejných sekvencí získaných pomocí odlišných sekvenačních platforem vůči sobě, poté došlo k vyhledávání genů obsažených v referenční sekvenci a na závěr byla provedena analýza kvality nalezených genů.

Poslední kapitola obsahuje shrnutí výsledků provedených hodnocení a metod porovnání bakteriálních genomů pocházejících z druhé a třetí generace sekvenačních platforem, které byly tématem této bakalářské práce.

1 Sekvenační technologie

První krok k pochopení role genetické informace učinili roku 1953 vědci Watson a Crick, oceněni Nobelovou cenou, když odhalili strukturu deoxyribonukleové kyseliny (DNA). Klíčem k porozumění genetické povahy organismů však stále zůstávalo dešifrování její sekvence. Od roku 1990 přicházely první návrhy na zmapování lidského genomu, jež se ukázaly být pouze začátkem moderní éry sekvenování DNA a vyústily v další invenci a zlepšení vývoje směrem k novým, pokročilým strategiím pro výkonné DNA sekvenování. Vývoj metod pro sekvenování DNA způsobil převrat v moderní biologii a zcela změnil naše chápání živých organismů. [1]

Sekvenátory se dnes dělí na 3 generace. Do první jsou zařazeni Maxam-Gilbert a Sangerovo sekvenování. Obě zmíněné techniky jsou časově a technicky náročné. V dnešní době se lze setkat jen se Sangerovým sekvenováním, které se využívá na menší projekty či dosekvenování určitých úseků díky jeho přesnosti. Druhá generace je začátkem rozvoje komerčních produktů. Dochází k paralelizaci sekvenování, a tím k jeho zrychlení a snížení ceny. Nejvyužívanějšími platformami jsou Roche, Illumina, SOLiD a Ion Torrent. Třetí generace se vyznačuje novým přístupem, umožňuje přímé sekvenování jediné molekuly DNA. Sem se řadí platformy od Helicos BioSciences, Pacific BioSciences nebo Oxford Nanopore Technologies (ONT). [2] Pro tuto bakalářskou práci jsou stěžejní platformy druhé a třetí generace.

1.1 Druhá generace sekvenátorů

Všechny platformy druhé generace jsou založeny na paralelním sekvenování velkého množství malých fragmentů DNA. Každá báze genomu je tak osekvenována několikrát. Posléze je potřeba zmapovat jednotlivá čtení a seřadit tyto fragmenty do posloupnosti celého genomu. Do druhé generace patří pyrosekvenování 454 Roche, které se odvíjí od luminiscenčního zachycení pyrofosfátové syntézy. Technologie 454 se v dnešní době již nevyužívá. Produkuje více než 700 párů bází dlouhá čtení, ale vyznačuje se chybovostí a vyšší cenou. Další je technologie SOLiD, která využívá sekvenování pomocí ligace. SOLiD produkuje čtení dlouhá v řádech desítek párů bází a je pomalejší než ostatní platformy druhé generace, proto se od ní upustilo. Jako další musí být zmíněny nejvyužívanější platformy od firmy Illumina, které bude věnována nadcházející část textu. Jako poslední stojí za zmínku platforma Ion Torrent, která využívá odlišného přístupu než předchozí platformy. Neodvíjí se od optických metod, ale zachycuje koncentraci vodíkových kationtů a měří tak pH. Díky tomu je tato platforma velmi levná a využívá se i v dnešní době. [3]

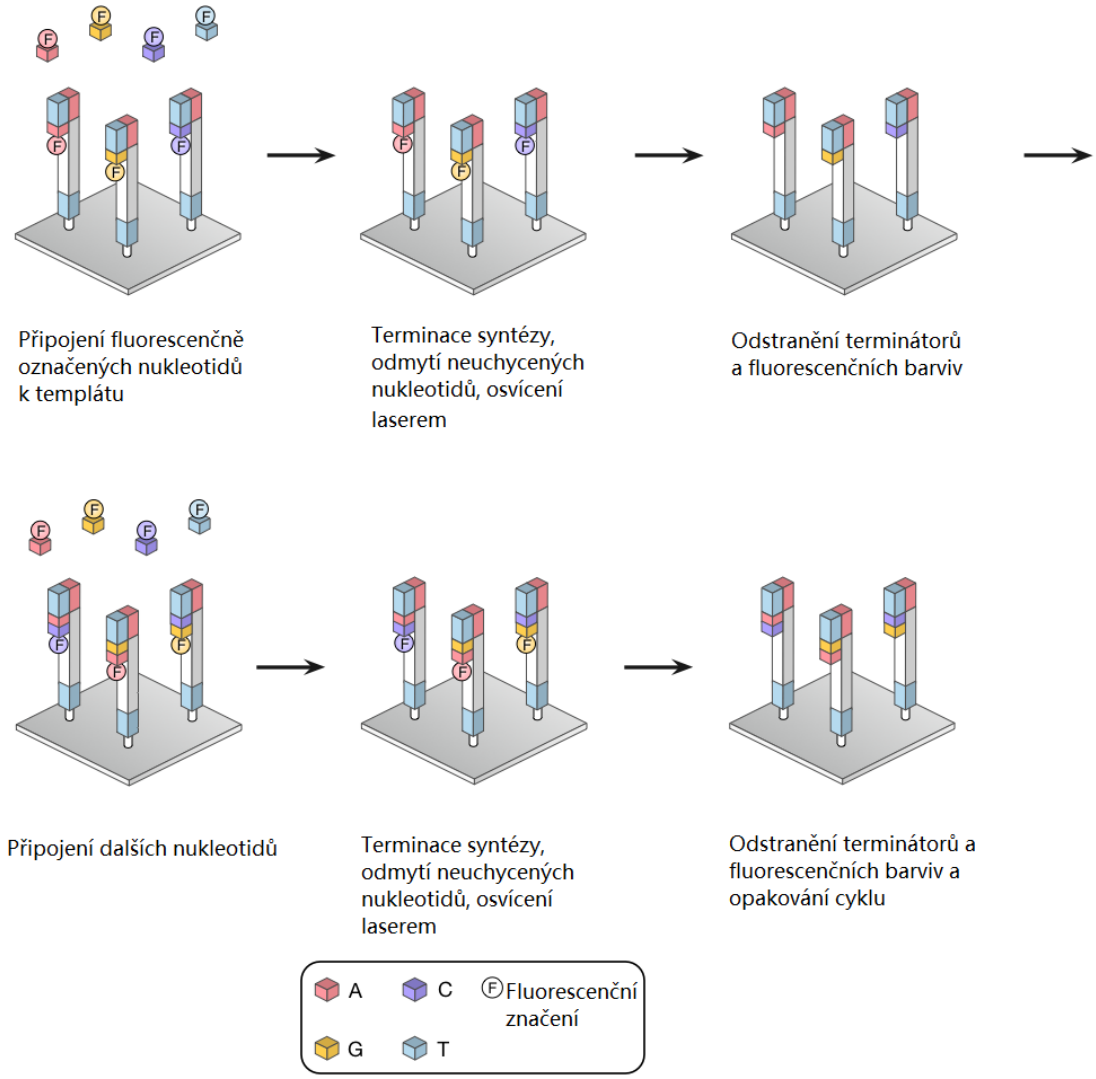
Druhou generaci lze využít pro osekvenování celého genomu nebo může být omezena na konkrétní oblasti zájmu. Sekvenování druhé generace se vyznačuje nízkou

chybovostí a poskytuje tak možnost zachycení variací a mutací ve struktuře DNA. [4]

1.1.1 Illumina MiSeq

Tato platforma patří do druhé generace sekvenačních technologií. Využívá princip sekvenace syntézou a zároveň reverzibilní terminátorovou reakci. Pro možnost sekvenace je zásadní příprava knihovny. DNA se nafragmentuje, dojde k ligaci s oligonukleotidovými adaptéry a následně k denaturaci. Vzniknou krátké úseky jednovláknové DNA. Takto připravená knihovna se nanese na destičku, na níž jsou napojeny adaptéry komplementární k těm, které ligovaly s fragmenty DNA. Tyto komplementární adaptéry hybridizují a následně slouží jako primery k syntéze řetězce DNA dle templátu. Po syntéze dochází k odstranění původní molekuly DNA a na destičce zůstanou pouze vzniklé kopie, které pomocí můstkové polymerázové řetězové reakce (PCR) vytvoří shluky jednovláknové DNA. Posledním krokem přípravy k sekvenování je odstranění reverzních vláken a přidání primerů, DNA polymerázy, čtyř typů nukleotidů (dNTP) s rozdílným barevným značením a reverzibilního terminátoru. [5]

Samotná sekvenace začíná navázáním DNA polymerázy k primeru templátu. Poté dochází k přiřazení komplementárního, fluorescenčně značeného dNTP k templátu. Důležitým krokem je následná terminace syntézy, při níž dochází k odmytí neuchycených nukleotidů. Díky tomuto kroku se sekvenování pomocí Illumina MiSeq označuje jako reverzibilní terminátorová sekvenace. Následně je destička osvícena dvěma typy laserů (pro A/C a G/T) a zachycuje se odpověď v podobě barevné fluorescence nukleotidů. Každý nukleotid vyzařuje jinou barvu, tak dochází k dekódování jednotlivých nukleotidů. Po zaznamenání odpovědi dojde k odstranění terminátorů a fluorescenčních barviv. V dalším cyklu se opět přidají nové primery, DNA polymerázy, barevně označené nukleotidy a reverzibilní terminátory. Následně je znovu zahájena syntéza, která je po přiřazení jednoho komplementárního nukleotidu terminátorem zastavena a postup se opakuje až po přečtení celého fragmentu. Nasyntetizované vlákno se odstraní, dopředné vlákno, které sloužilo jako templát, se převede na reverzní vlákno a celý proces probíhá od začátku. [6] Postup sekvenace je zobrazen na obrázku 1.1.



Obr. 1.1: Postup terminátorové reverzibilní sekvenace. Převzato z [7].

Při tomto typu sekvenování může docházet k častým chybám. Jedním z důvodů může být zmíněná reverzibilní terminace syntézy. Při nedostatečně důsledné terminaci může v některých pozicích dojít k uchycení více než jednoho nukleotidu ve stejném cyklu. Na druhou stranu může nastat opačný problém, kdy v některém z cyklů nedojde k reverzibilní terminaci, ale terminace se stane trvalou. Tento scénář vede k postupnému úbytku míry signálu. Dalším důvodem k chybovosti mohou být záměny mezi A/C a G/T i přes použití čtyř rozdílných fluorescenčních značení. Platforma Miseq excituje nukleotidy dvěma lasery tak, že značky pro A/C a G/T jsou excitovány vždy stejným laserem a vykazují signál o podobném spektru. Nárůst chyb může způsobit i úbytek nebo úplné vyčerpání DNA polymerázy a postupné vysvěcování fluorescenčních značek. [8]

Illumina Miseq produkuje velké množství poměrně krátkých čtení (milióny čtení o délce v řádech stovek párů bází). Doba sekvenace se pohybuje v rozmezí od 2 do

24 hodin při generaci 25 milionů 300 bp dlouhých pair-end čtení. I přes nevýhody této platformy je hlavně díky své nízké ceně při přepočtu na jednotlivé báze jednou z nejhojněji používaných sekvenačních technologií. [9]

1.2 Třetí generace sekvenátorů

Platformy třetí generace se vyznačují produkcí dlouhých čtení, rychlostí sekvenování a snadnou přípravou knihovny. Tyto platformy jsou zároveň nejnovější na trhu, a proto stále dochází k jejich vylepšování a rozvoji. Patří mezi ně sekvenátory od společnosti PacBio, které jsou založeny na principu sekvenování jedné molekuly v reálném čase (z angl. single molecule real time sequencing neboli SMRT) a fluorescenčním snímání. Jejich výhodou je produkce dlouhých čtení a snaha o zvýšení přesnosti, nevýhoda spočívá ve vysoké ceně vybavení. Další platforma spadající do třetí generace je založena na sekvenování nanopórem a bude jí věnována následující část textu.

1.2.1 Oxford Nanopore Technologies MinION

Zařízení MinION je založené na technologii nanopórů, které jsou uloženy v membráně ze syntetického polymeru s velkým elektrickým odporem. Samotný nanopór je tvořen alpha-hemolyzinem kovalentně připojeným k molekule cyklodextrinu, která představuje vazebné místo pro nukleotidy. Sekvenační komůrka (z angl. flowcell) sekvenátoru MinION, kam se vkládají vzorky, obsahuje 512 kanálů, díky kterým je možné paralelně sekvenovat až 512 jednotlivých DNA molekul. Výkon jednotlivých kanálů se z hlediska počtu produkovaných čtení liší, protože některé póry jsou aktivnější než jiné. [10]

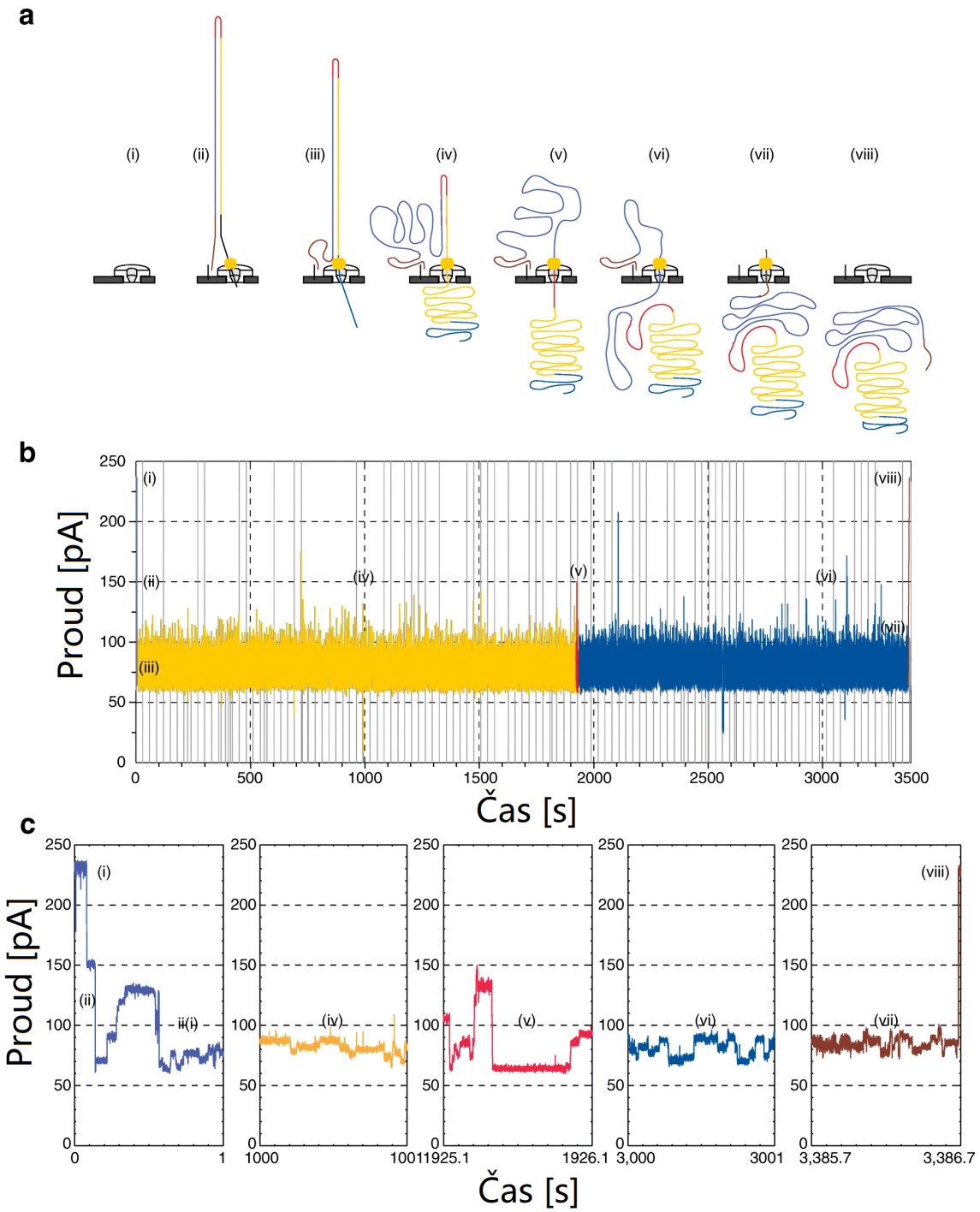
Stejně jako u sekvenačních technologií druhé generace, je zapotřebí příprava knihovny. Jelikož MinION dokáže zpracovat obě vlákna šroubovice, je nejvhodnější použít dvouvláknovou DNA pro dosažení nejvyšší kvality čtení. DNA se napřed nastříhá na menší fragmenty, upraví se poškozená DNA a vytvoří tupé konce, na které se naváže adeninový nukleotid. Poté dojde k ligaci adaptérů. Knihovna využívá dva druhy adaptérů, vedoucí a vlásenkový, které se připojí na opačné konce řetězce.

Sekvenování začíná na 5' konci vlákna s vedoucím adaptérem, kdy motorový protein nanopóru rozezná tento adaptér a rozvolní vlákna DNA. Nanopórem tak může projít templátové vlákno rychlostí, kterou určí motorový protein. Poté na řadu přichází vlásenkový adaptér a komplementární vlákno. Vlásenkový adaptér spojuje templátové a komplementární vlákno, čímž je umožněno dvouvláknové čtení. Průběh je znázorněn na obrázku 1.2 a).

Na membránu s nanopóry je přivedeno napětí a dochází k průchodu iontových proudů. Když molekula DNA prochází skrz nanopór, dochází ke změnám těchto proudů. Každá báze nukleotidů molekuly DNA vyvolá jinou změnu proudu. Tyto změny můžeme zaznamenávat v čase a dekódovat tak jednotlivé nukleotidy. Příklad záznamu lze vidět na obrázku 1.2 b). Za účelem snížení šumu jsou měření proudu zpracována a převedena na posloupnost událostí do tzv. "squiggle plot"(např. 1.2 c)). Data jsou vyhodnocována na speciálním ASIC mikročipu. [11]

Basecalling, neboli rozkódování signálu na jednotlivé báze, neprobíhá na úrovni jednoho nukleotidu, ale vždy je snímán signál po pěticích bází.

Výhodou přístroje MinION jsou velmi malé rozměry, velikostí jde vlastně o kapesní zařízení, které se připojuje k počítači prostřednictvím USB a lze jej využít při práci v terénu. Sekvenování celé molekuly pomocí nanopórů odstraňuje potřebu amplifikace pomocí PCR nebo chemického značení vzorku, a je tedy menší pravděpodobnost, že dojde k vytvoření určitého artefaktu či poškození zkoumané DNA. Přístroj je navíc schopen číst až desítky tisíc bází dlouhé sekvence a to v reálném čase. Naopak jeho největší nevýhodou je poměrně značná chybovost, která dosahuje až 10 %. Je dána především dekódováním signálu po 5 nukleotidů dlouhých celcích. [12]



Obr. 1.2: Sekvenování nanopórem. a) Průchod vlákna DNA nanopórem: (i) Nanopór na povrchu membrány. (ii) Vedoucí adaptér vlákna začíná prostupovat nanopórem. (iii) Poté prochází templátové vlákno. (iv) Následuje vlásenkový adaptér. (v) Jako poslední nanopórem prochází vlákno komplementární. (vi) Ukončení sekvenace. (viii) Nanopór je opět volný. b) Záznam proudu procházejícího nanopórem v čase. c) Konkrétní příklady událostí proudu. Převzato z [16].

2 Sestavování genomu

Dnešní sekvenátory produkují velké množství čtení. Pro sestavení celého genomu je potřeba tyto kousky poskládat a seřadit. Sestavení genomu (z angl. assembly) popisuje jak samotný postup, tak výsledek sestavených dat. Důležitým parametrem sestavení je předpoklad, že každá báze genomu musí být pokryta dostatečným počtem čtení. Tak je zaveden parametr pokrytí (z angl. coverage). Sestavení lze provést na základě referenčního genomu nebo se přistupuje k tzv. *de novo* sestavení. Rozhodnutí závisí na následném využití sestavených genomů a časové náročnosti sestavení. K samotnému procesu sestavení genomu lze také využít několik různých typů algoritmů, mezi které patří např. de Bruijnův graf (DBG) nebo Overlap Layout Consensus (OLC).

2.1 De Bruijnův graf (DBG)

Pro většinu sekvenátorů, které produkují velké možství krátkých čtení (např. Illumina Miseq) je nejvhodnější použít algoritmy z oblasti teorie grafů, tzv. de Bruijnovy grafy. Algoritmus DBG nevyužívá celých čtení, ale rozděluje je na k-mery (krátké úseky bází) předem definované délky.

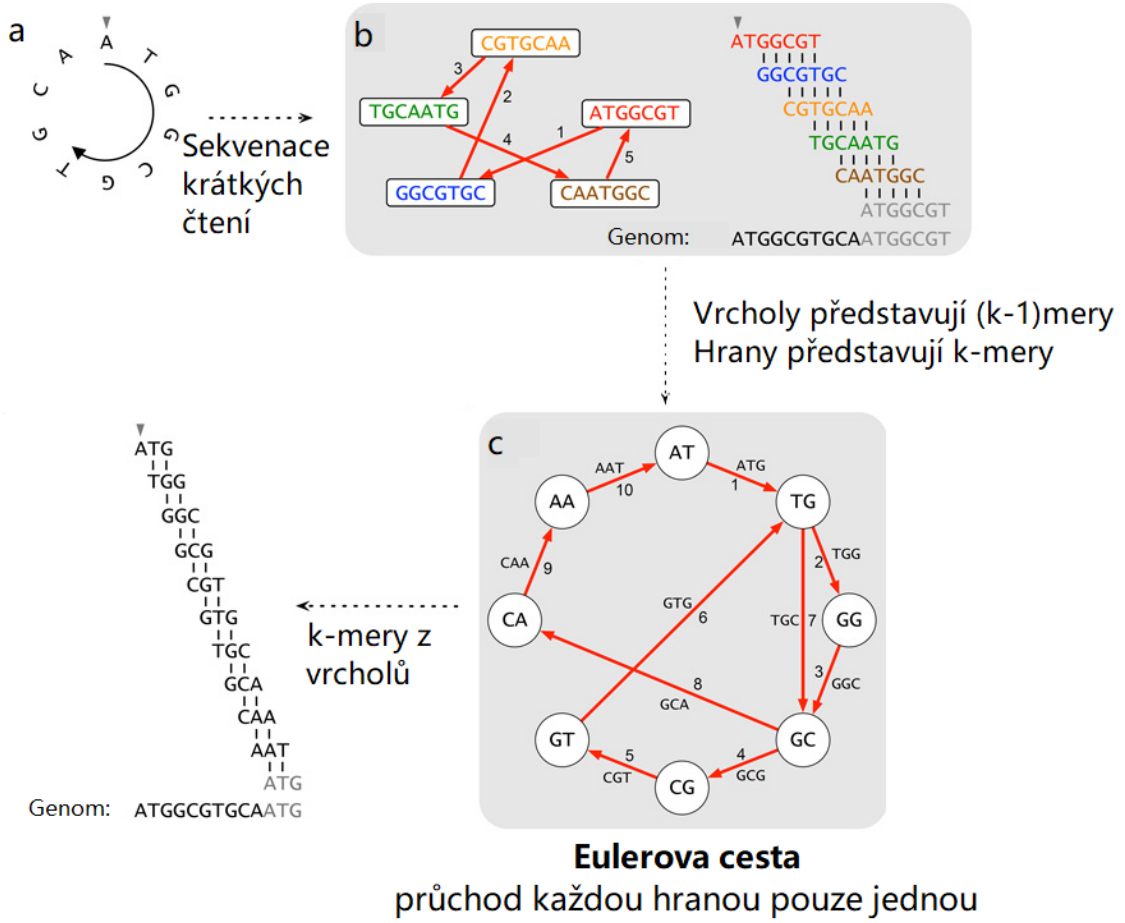
Než algoritmus začne se samotným sestavováním genomu, je potřeba vstupní čtení filtrovat a opravit případné chyby vzniklé sekvenováním. Následně se přechází k tvorbě de Bruijnova grafu. Na vrcholy grafu se dosadí prefixy a sufixy k-mer, které se získají zkrácením původních k-mer na jejich koncích o jednu bázi. Prefix představuje k-mer bez poslední báze a sufix k-mer bez první báze. Poté se vykreslí hrany ve formě k-mer spojující vrcholy na základě toho, zda mají společný konkrétní prefix a sufix. Hranám se také přidá směr pro určení cesty. Směr hran mezi vrcholy naznačuje, že k-mery na těchto vrcholech se vyskytují postupně v jednom nebo více čteních. Algoritmus prochází každou hranou pouze jedenkrát. Postup je znázorněn na obrázku 2.1. Tomuto principu se říká Eulerova cesta, díky které se jednotlivé k-mery spojují do větších souvislých posloupností tzv. kontigů. Ty se spojují do výsledné sekvence. Jednoznačné úseky sekvence tvoří v de Bruijnově grafu nerozvětvené cesty, které usnadňují celkové sestavování sekvence. Naopak kontigy tvořící rozvětvené cesty tedy kontigy, jejichž k-mery jsou obsaženy v jiných kontizích, nepřidávají nic do výsledné sekvence, zvyšují však pokrytí úseku. [13], [14]

Na rozdíl od algoritmů využívajících překryvu šetří DBG paměť algoritmu pro sestavování genomu a zároveň zkracuje čas zpracovávání, protože k zachycení překryvů slouží pouze graf a nejsou tedy explicitně vypočítávány.

Algoritmus DBG komplikují chyby sekvenování, ale mnoho takových chyb lze snadno rozpoznat podle jejich struktury v grafu. Programy pro sestavování genomu

využívající DBG dokáží takové struktury vyhledat a odstranit chybové vrcholy a vrcholy s nízkým pokrytím. Hlavní nevýhodou DBG je ztráta informací způsobená rozkladem jednotlivých čtení na cestu k-merů. De Bruijnovy grafy vytvářejí více vrcholů pro každé čtení a tyto vrcholy nemusí po přidání hran z jiných čtení vytvářet lineární cestu, což znamená, že v grafu se mohou objevit cesty tvořící sekvenci, která není podporována původními čteními.

Algoritmus de Bruijnova grafu se používá převážně pro sestavení bakteriálních genomů. Sestavení lidského genomu by pro něj představovalo velkou paměťovou náročnost, a proto se k těmto účelům raději nevyužívá. [15]



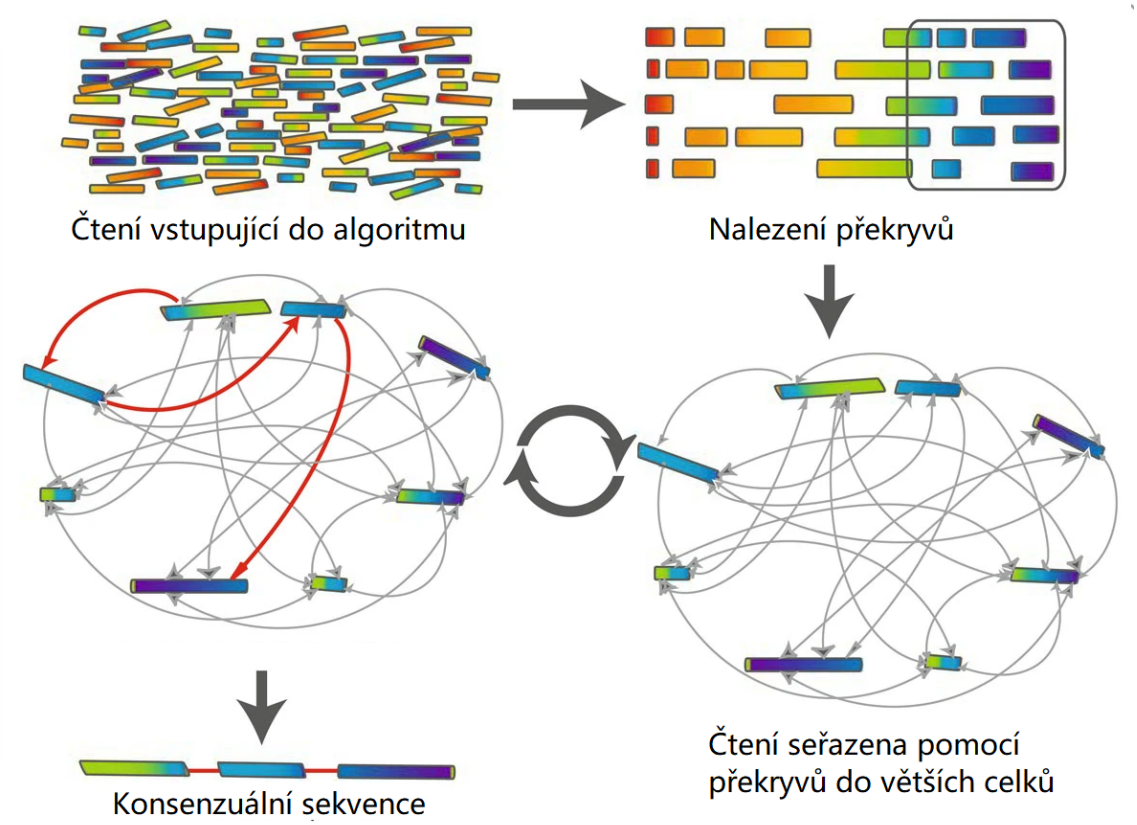
Obr. 2.1: Princip tvorby de Bruijnova grafu. a) Příklad krátké kruhové molekuly DNA. b) Sangerův sekvenační algoritmus - čtení reprezentována jako vrcholy v grafu, hrany představují zarovnání mezi čteními. c) Sestavení sekvence na základě Eulerovy cesty, kde k = 3. Převzato z [13].

2.2 Overlap Layout Consensus (OLC)

Algoritmus OLC byl původně používán pro dlouhá čtení Sangerovy sekvenační metody. S příchodem další generace sekvenátorů se do popředí dostaly algoritmy sestavení pro krátká čtení založené na de Bruijnově grafu. Avšak s rozvojem třetí generace a produkcí dlouhých čtení, například z platformy Oxford Nanopore Technologies MinION, a jejich větší chybovostí se použití DBG stává nevhodným. Nanopórové sekvenování tak pomáhá návratu sestavování genomu založeném na principu překryvu. [16]

Jelikož byly OLC metody původně vyvinuty pro Sangerovo sekvenování, které se vyznačuje nižší chybovostí, je potřeba opravit sekvenační chyby nanopórové sekvenace ještě před začátkem sestavování genomu. K tomu slouží předzpracování, které je časově velmi náročné, jelikož zahrnuje mapování všech čtení. Jak název napovídá, metoda hledá překryvy jednotlivých čtení. Princip hledání překryvu spočívá v zarovnání dvou čtení a nalezení jejich nejdelší společné části. Čtení se na základě překryvů skládají do kontigů. Teoreticky by tímto postupem mělo být možno vytvářet dlouhé kontigy z původních čtení, pokud má oblast dostatečné pokrytí. Ve skutečnosti se však genom skládá z dlouhých repetitivních úseků, u kterých je těžké provést sestavení. Poté, co byly fragmenty čtení seřazeny do kontigů, je výsledkem graf, který do vrcholů umístí jednotlivé kontigy a spojí je, když je jejich překryv větší než mezní hodnota. Tak jsou kontigy spojeny do větších celků. K sestavení celé sekvence je zapotřebí vícenásobného zarovnání těchto celků. [17], [18] Postup metody OLC je znázorněn na obrázku 2.2.

Jak již bylo zmíněno výše, některá překrývající se čtení nemusí být identická kvůli chybám sekvenování nebo polymorfismu. Tyto konflikty lze vyřešit, vezme-li se v úvahu skóre kvality pro každou bázi. U platformy MinION lze finální úpravu sestavení sekvence provést pomocí softwaru Nanopolish, který zlepšuje kvalitu řazení bází přehodnocením a maximalizací pravděpodobností pro každou z nich dle událostí iontových proudů přístupných v souborech FAST5. [19]



Obr. 2.2: Postup metody OLC pro sestavení sekvence genomu. Převzato z [18].

3 Formáty a struktura zápisu sekvenačních dat

Z Fakultní nemocnice Brno byla poskytnuta data představující 6 osekvenovaných genomů bakterie *Klebsiella pneumoniae*. Genomy byly osekvenovány na 2 platformách: Illumina Miseq a Oxford Nanopore Technologies MinION. V této kapitole jsou uvedeny nejčastěji využívané formáty, se kterými se lze setkat při práci se sekvenačními daty.

3.1 FASTA

FASTA je textový formát sloužící k zápisu genomických a proteomických sekvencí. Skládá se z hlavičky, která je uvozená znakem „>“ a kde se nachází informace o dané sekvenci. Druhou částí je samotná posloupnost nukleotidů nebo aminokyselin zapsaná dle IUPAC nomenklatury.

3.2 FASTQ

Tento formát vychází z formátu FASTA, ale oproti němu se vyznačuje informací o kvalitě přečtení jednotlivých nukleotidů. Stejně jako FASTA obsahuje na prvním řádku hlavičku, tentokrát uvozenou znakem „@“, a na druhém řádku samotnou sekvenci. Třetí řádek obsahuje pouze znak „+“. Na čtvrtém řádku se nachází znaky zmíněné kvality přečtení báze, přičemž každý znak odpovídá jedné dané bázi ze sekvence dle jejich pořadí. Znaky specifikující kvalitu pocházejí ze základního kódování ASCII. [20]

3.3 FAST5

Formát FAST5 slouží především k ukládání výstupu ze sekvenátorů společnosti Oxford Nanopore Technologies. Vychází z formátu HDF5, což je velmi flexibilní datový model, knihovna a formát souborů pro ukládání a správu dat. Je schopen ukládat neomezené množství datových typů. Formát FAST5 kromě samotných proudových hodnot, které jsou získány měřením na nanopórech, ukládá také informace o samotném průběhu sekvenace, jako je například doba jejího trvání, počet přečtených čtení, délka jednotlivých čtení, informace o kvalitě přečtení bází a to pro každý kanál nacházející se v sekvenační komůrce sekvenátoru.

3.4 SAM/BAM

Formáty SAM a BAM slouží k zápisu zarovnání čtení produkovaných sekvenátory k referenci. Ve formátu SAM se nachází hlavička, uvozena znakem „@“, a poté samotné zarovnání. Každý řádek zarovnání obsahuje 11 sloupců, ve kterých se nachází základní informace o zarovnání, jako je název zarovnaného úseku a referenční sekvence, parametr zvaný FLAG, který představuje součet bitových příznaků daného čtení nebo vlákno zvané CIGAR, ve kterém jsou obsaženy rozdíly bází mezi čtením a referencí. Dále se v zarovnání nachází informace o pozici mapování, kvalitě mapování a kvalitě přečtení báze. Tyto parametry jsou reprezentovány svou hodnotou nebo 0 či znakem „*“, pokud daná hodnota chybí. Kromě zmíněných 11 sloupců, které musí obsahovat každý řádek zarovnání ve formátu SAM, lze také připojit volitelný počet sloupců se specifickými informacemi pro dané zarovnání. Formát BAM je binární reprezentací SAM a obsahuje také stejnou informaci, s tím rozdílem, že je komprimována. [21]

4 Sestavení poskytnutých genomů

Jak bylo zmíněno výše, genomy poskytnuté z FN Brno byly osekvenovány na 2 platformách: Illumina Miseq a Oxford Nanopore Technologies MinION. Zpracování dat z Illumina Miseq proběhlo v prvé řadě pomocí odstřižení adaptérů, které byly ke čtením připojeny při sekvenaci. Následovalo mapování čtení k referenční sekvenci, filtrace nenamapovaných a nevhodně namapovaných čtení a vytvoření konsenzuální sekvence z daného zarovnání pomocí příslušné sady nástrojů, zatímco data ze sekvenátoru MinION musela být napřed převedena do znakové formy, demultiplexována a až poté sestavena do podoby kompletních genomů.

4.1 Sestavení genomů Illumina Miseq

Na platformě Illumina Miseq proběhlo paired-end sekvenování 6 genomů bakterie *Klebsiella pneumoniae*. Poskytnutá data představují dva soubory ve formátu FASTQ pro každý genom.

Tato data byla v prvé řadě zbavena adaptérů, které obdržela při přípravě knihovny na sekvenaci. Odstřižení těchto adaptérů bylo provedeno pomocí nástrojů Trimmomatic (verze 0.39, [22]) za použití knihovny adaptérů TruSeq3-PE.

Následně byla data zarovnána pomocí softwaru Burrows–Wheeler Alignment (BWA verze 0.7.17, [23]). Zmíněný software se využívá k zarovnání čtení vůči referenčnímu genomu a obsahuje 3 algoritmy. Pro poskytnutá data byl vybrán algoritmus BWA-MEM, který je vhodný pro 70 bp - 1 Mbp dlouhá čtení a vyznačuje se přesností a rychlostí zarovnání. Genomy byly tedy pomocí tohoto algoritmu zarovnány vůči referenčnímu genomu NC_012731.1 získanému z RefSeq databáze NCBI. Výstupem algoritmu je zarovnání ve formátu SAM.

Pomocí sady nástrojů SAMtools (verze 1.10, [21]) bylo zarovnání každého genomu převedeno na formát BAM, uspořádáno a poté byla všechna jeho nenamapovaná čtení a čtení, která se naopak namapovala více jak dvakrát, vyfiltrována.

V rámci posledního kroku byla pomocí sady nástrojů BCFtools (verze 1.10.2, [24]) ze zarovnání vygenerována konsenzuální sekvence ve formátu FASTQ, která byla následně převedena na formát FASTA. Takto byla obdržena kompletní sekvence genomů sekvenovaných na platformě Miseq.

4.2 Sestavení genomů Oxford Nanopore Technologies MinION

Data z platformy MinION představují 6 genomů bakterie *Klebsiella pneumoniae*, které byly osekvenovány v jednom běhu využívající flowcell FLO-MIN107 a soupravu pro přípravu knihovny (z angl. kit) s označením SQK-RBK004 (Rapid Barcoding Kit). Data byla obdržena ve formátu FAST5, který obsahuje hodnoty naměřeného proudu procházejícího nanopóry. V prvním kroku musely být tedy tyto hodnoty převedeny do znakové podoby (z angl. basecalling). Převedení bylo provedeno pomocí programu Guppy (verze 3.6.0, [25]), který k urychlení svých výpočtů využívá grafický procesor. Při sekvenování byly fragmenty DNA každého z šesti genomů označeny rozdílnými identifikátory (z angl. barcode). Tyto identifikátory představují umělou DNA, která se v původním vzorku nenachází, tudíž byly v následujícím kroku odstřiženy a jednotlivá čtení byla rozdělena do souborů odpovídajících vzorkům. Tento postup se nazývá demultiplexace a byla též provedena pomocí programu Guppy. [26] Poté bylo provedeno zarovnání čtení pomocí algoritmu Flye (verze 2.8, [27]), který na rozdíl od jiných algoritmů pro sestavení sekvencí generuje disjunktní kontigy (z angl. disjointigs). Takto vytvoří řetezec disjunktních segmentů, ze kterého sestaví graf. Pomocí grafu dojde k sestavení výsledné sekvence nukleotidů bez přispění podkladu referenčního genomu neboli tzv. *de novo*. Výstupem programu je soubor ve formátu FASTA s výslednou sekvencí.

5 Hodnocení kvality genomů

5.1 Hodnocení kvality sekvenace genomů

Před sestavením kompletních sekvencí byla u genomů otestována kvalita jejich přečtení při sekvenování. Toto hodnocení bylo provedeno na poskytnutých datech z Illumina Miseq a na textovém souboru shrnujícím průběh sekvenace na platformě MinION.

5.1.1 Hodnocení kvality sekvenace dat z Illumina Miseq

Data sekvenovaná pomocí Miseq byla analyzována v programu FastQC (verze 0.11.5, [28]). Vstupní soubory byly dodány ve formátu FASTQ, který kromě samotné sekvence obsahuje i kvalitu jednotlivých bází ve formě zakódovaného Phred skóre Q. Kvalita analyzovaných sekvencí pocházejících z Illumina Miseq bude znázorněna pomocí 4 grafů, které charakterizují genom EB359 (Obr 5.1 až Obr 5.4). Kompletní sady grafů pro všechny genomy jsou umístěny v příloze A na Obr. A.1 až Obr. A.4.

První sada A.1 představuje krabicové grafy ukazující seskupení skóre kvality v každé pozici v rámci jednotlivých čtení genomů. Osa x uvádí pozice v rámci čtení v jednotkách bp tak, že prvních 10 pozic je uvedeno samostatně, dále jsou pozice seskupeny do oken o určité šířce závisející na délce čtení. Osa y uvádí hodnoty skóre kvality (Phred skóre) v rozsahu 0 až 40. Modrá křivka protínající každý z grafů představuje průměrnou hodnotu Phred skóre pro každou pozici napříč čteními. Naopak červená čára, která je obsažena v každém boxu, odpovídá mediánu Phred skóre pro danou pozici. Z grafů je zřejmé, že jak průměrné skóre kvality, tak jeho medián u všech čtení genomů nejprve vzroste a poté začíná klesat až dosáhne své nejnižší hodnoty na poslední pozici ve čtení. Tento pokles je často způsoben útlumem signálu nebo fázováním genomu během sekvenačního běhu. Nejkvalitněji nasekvenovaných čtení dle těchto grafů dosahuje genom EB359 5.1, jehož průměrná hodnota Phred skóre na všech pozicích čtení je vyšší než Q = 28 a také nejnižší hodnota jeho mediánu odpovídá Q = 32. Naopak genom KP268 A.1a, s nejnižší hodnotou průměrného Phred skóre Q = 24 a nejnižší hodnotou mediánu tohoto skóre Q = 21, a genom EB360 A.1b, s nejnižší hodnotou průměrného Phred skóre Q = 23 a nejnižší hodnotou mediánu tohoto skóre Q = 19, obsahují nejméně kvalitně osekvenovaná čtení.

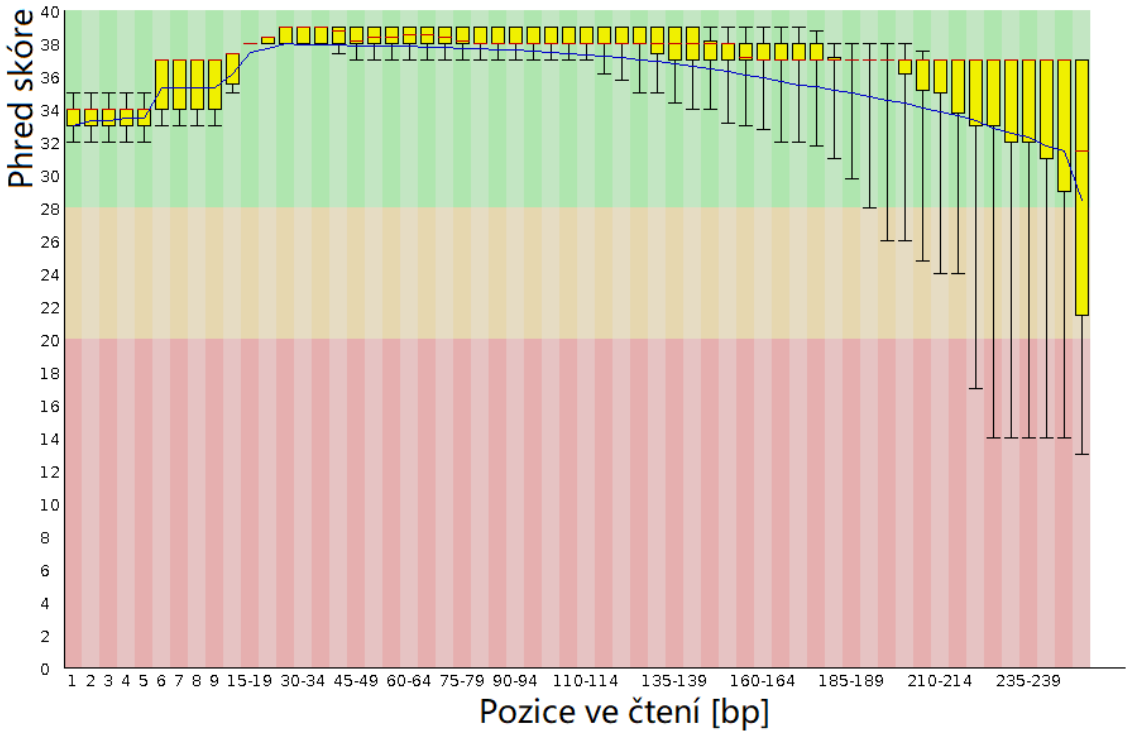
Druhá sada grafů A.2 vyjadřuje distribuci hodnoty Phred skóre na celkový počet čtení jednotlivých sekvencí genomů. Na ose x se nachází průměrné hodnoty Phred skóre a na ose y počet čtení nasekvenovaných pro daný genom. Grafy všech genomů vykazují vysokou hodnotu Phred skóre napříč čteními a tudíž vysokou kvalitu čtení sekvenátoru. Grafy všech těchto sekvencí začínají vzrůstat nad hranici 100 000 čtení

kolem hodnoty Phred skóre Q = 30. Pouze graf genomu EB360 (Obr. A.2b) začíná vzrůstat nad hranici 100 000 čtení v okolí hodnoty Phred skóre Q = 22. To znamená, že tato sekvence obsahuje větší množství méně kvalitních čtení. Zároveň z grafů vyplývá, že u všech genomů se kvalita většiny čtení pohybuje kolem hodnoty Q = 36, což značí, že přesnost sekvenace dosahuje až 99,99 %.

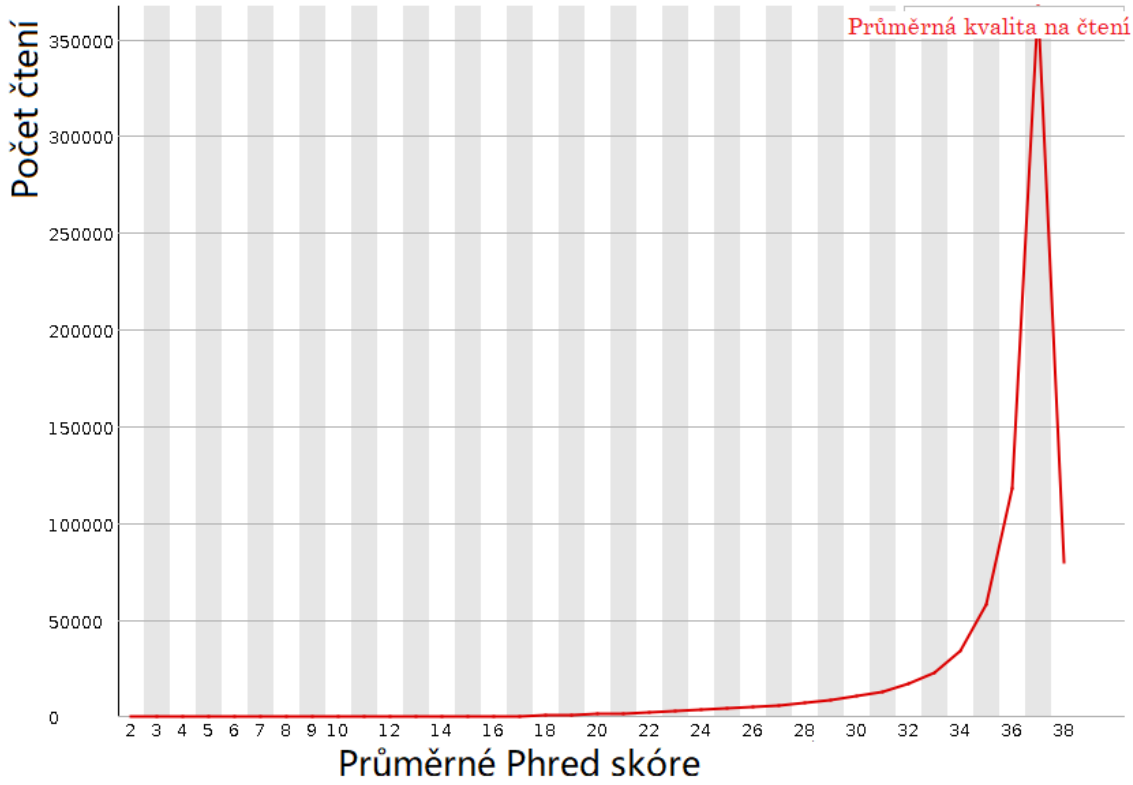
Třetí sada A.3 představuje procentuální zastoupení jednotlivých nukleotidů ve čteních. Na ose x se nachází pozice ve čteních stejně jako u první sady grafů a na ose y zastoupení nukleotidů v procentech. V ideálním případě by daný graf po sekvenaci DNA měl mít konstantní průběh procentuálního zastoupení pro každý nukleotid tak, že tyto průběhy pro nukleotidy, jejichž báze jsou komplementární dle pravidel pro párování, dosahují stejných hodnot. Uvedené grafy se narozdíl od ideálního průběhu vyznačují fluktuací v prvních a posledních pozicích v rámci čtení. Zastoupení nukleotidů s bázemi C a G převyšuje ve všech grafech 25 % a zastoupení nukleotidů obsahující báze A a T 20 %.

Čtvrtá sada grafů A.4 obsahuje informaci o obsahu adaptérů, které byly k jednotlivým čtením připojeny při přípravě knihovny na sekvenování. Osa x udává pozici ve čteních stejně jako u první a třetí sady grafů, osa y určuje procentuální zastoupení adaptérů. Z uvedených grafů je patrné, že čtení genomů EB359 5.4, KP268 A.4a a EB360 A.4b neobsahují žádné z uvedených adaptérů a je tedy zřejmé, že adaptéry byly odstraněny již v rámci sekvenace, zatímco ze čtení genomů KP1278 A.4c, KP1174 A.4d a KP1268 A.4e nebyly odstraněny všechny adaptéry, a tak jejich grafy na místě posledních pozic dosahují 10 - 15 % zastoupení adaptérů. Tyto adaptéry byly následně odstraněny v rámci filtrace před sestavením kompletních sekvencí genomů.

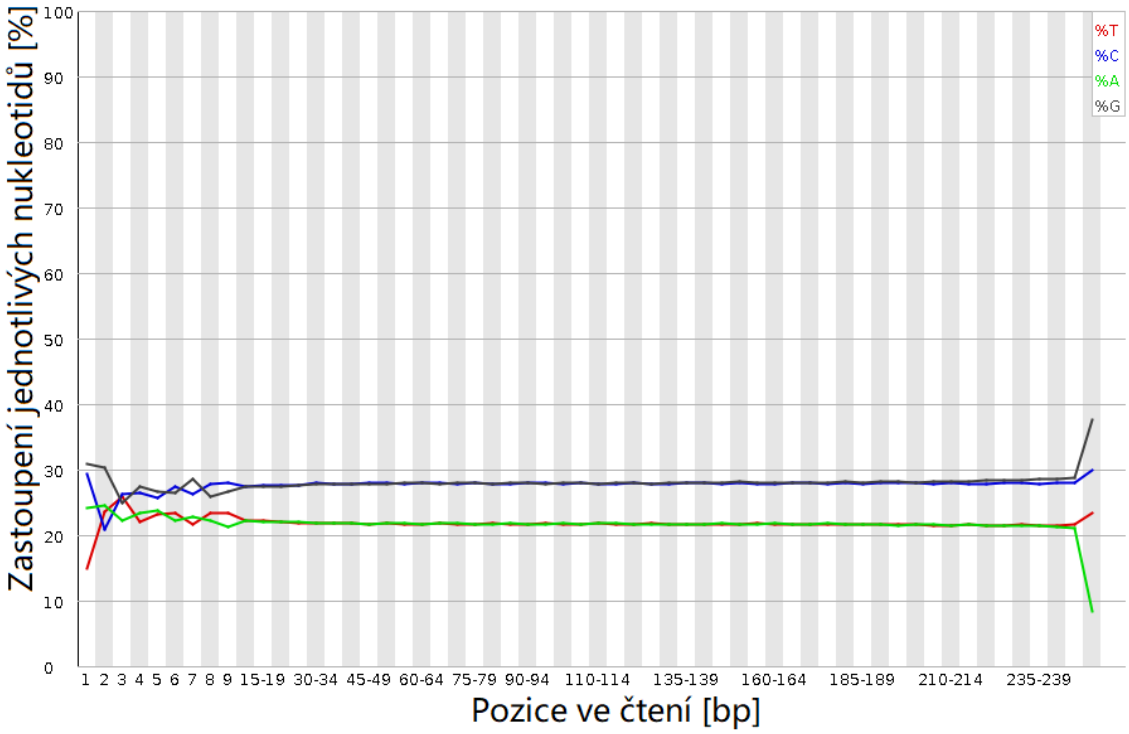
Výstup programu FastQC obsahuje další informace obsažené v grafech v příloze A. Jedná se o grafy znázorňující stupeň duplikace v genomech (Obr. A.5), počet k-mer v genomech (Obr. A.6), obsah nejednoznačných nukleotidů v genomech (Obr. A.7), průměrný obsah bází G a C v genomech (Obr. A.8), kvalitu sekvenace na pozici v rámci destičky sekvenátoru (Obr. A.9) a distribuci délky sekvence v genomech (Obr. A.10).



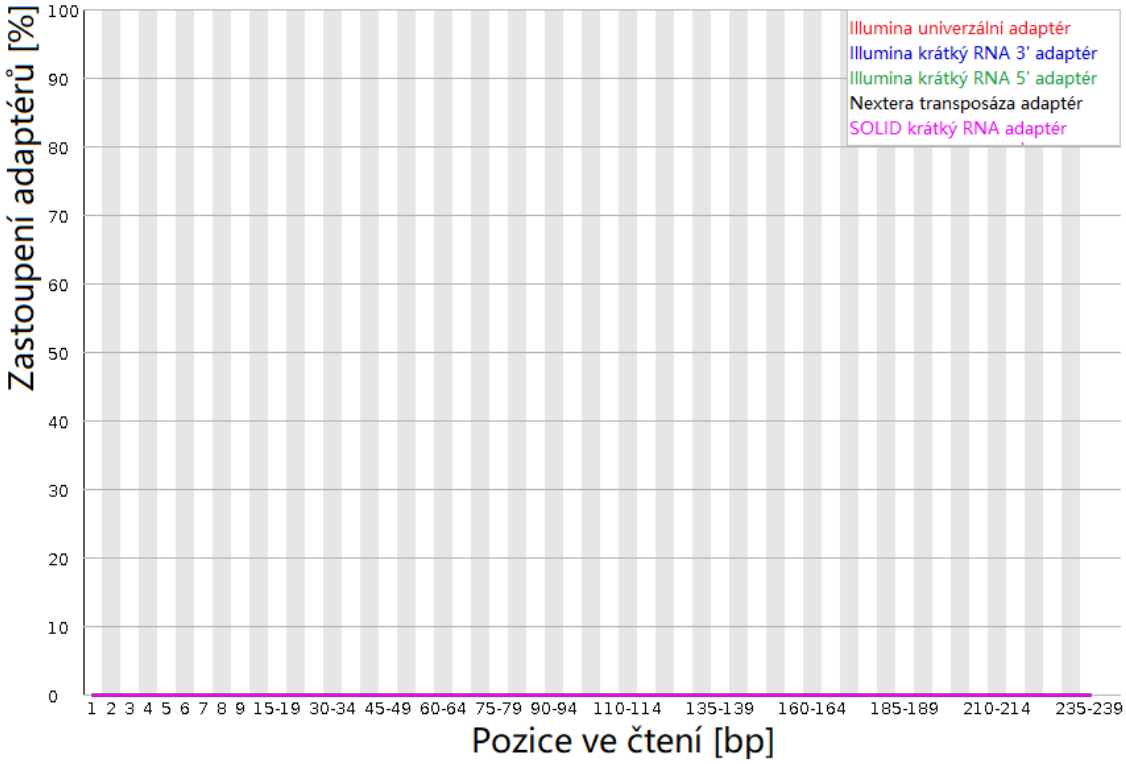
Obr. 5.1: Seskupení skóre kvality v každé pozici jednotlivých čtení genomu EB359 sekvenovaného pomocí Illumina Miseq.



Obr. 5.2: Distribuce Phred skóre na celkový počet čtení sekvence genomu EB359 sekvenovaného pomocí Illumina Miseq.



Obr. 5.3: Procentuální zastoupení přečtených bází pro každý ze čtyř nukleotidů v každé pozici jednotlivých čtení genomu EB359 sekvenovaného pomocí Illumina Miseq.



Obr. 5.4: Procentuální zastoupení adaptérů v každé pozici jednotlivých čtení genomu EB359 sekvenovaného pomocí Illumina Miseq.

5.1.2 Hodnocení kvality sekvenace dat z ONT MinION

Kvalita sekvenace dat získaných pomocí platformy MinION byla analyzována v programu MinIONQC (verze 1.4.1, [29]). Jako vstup byl použit textový soubor obsahující kompletní shrnutí průběhu sekvenace. Kvalita osekvenovaných dat pocházejících z platformy MinION je znázorněna pomocí souboru grafů od Obr. 5.5 až po Obr. 5.8.

První graf 5.5 představuje mapu sekvenační komůrky s 512 kanály a dílčími grafy vyjadřujícími průběh sekvenace pro každý kanál. Z grafu je patrné, že zdaleka nebyly využity všechny kanály, tzn. že byl využit jen omezený počet nanopórů. Zároveň je z barevného rozlišení grafu zřejmé, že kanály převážně produkovaly kvalitní čtení, která odpovídají zelenožlutému zbarvení. Modrofialové zbarvení označuje čtení s kvalitou nižší než Q = 7. Přestože jsou v grafu také viditelná, jejich množství je výrazně menší.

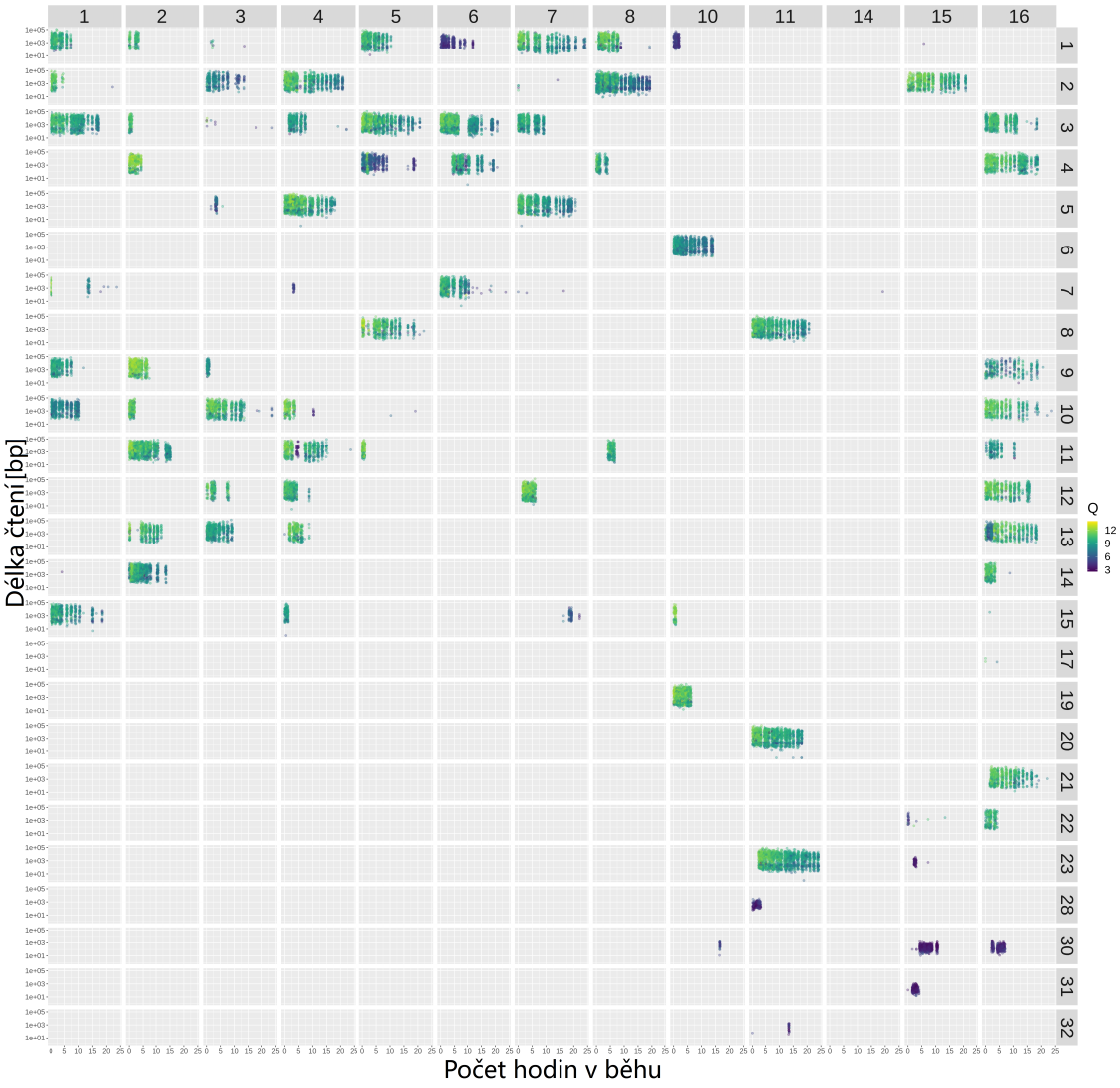
Obr. 5.6 vyjadřuje poměr průměrné hodnoty Phred skóre ve čtení vůči počtu takovýchto čtení. Obrázek je rozdělen na dva grafy. Vrchní graf pojednává o všech nasekvenovaných čteních, zatímco spodní graf zahrnuje pouze čtení s hodnotou Phred skóre vyšší nebo rovno Q = 7. Z grafů je patrné, že průměrná kvalita největšího počtu čtení se pohybuje kolem Q = 10. Nejvyšší hodnota Phred skóre, které v grafu dosahuje počet čtení pouze v řádech jednotek, odpovídá Q = 14. Podobné množství čtení obsahuje naopak nejnižší hodnotu uvedenou v grafu Q = 3.

Na Obr. 5.7 se nachází informace o počtu čtení určité délky. Osa x udává délku čtení v jednotkách párů bází v logaritmickém měřítku a osa y počet čtení. Obrázek je opět rozdělen na dva grafy tak, že vrchní graf obsahuje informace o všech čteních, zatímco spodní graf udává pouze délky a počty čtení, která obsahují průměrné skóre kvality vyšší nebo rovno Q = 7. Oba tyto grafy obsahují dva výrazné shluky s velkým počtem čtení. První skupina je seskupená v rozmezí 100 až 1000 bp a druhá skupina zabírá rozmezí 1000 až 10 000 bp. U spodního grafu, který obsahuje pouze kvalitní čtení, je první skupina s kratšími čteními zredukována oproti vrchnímu grafu, což naznačuje, že kratší čtení nedosahují takové kvality jako delší čtení. Nejvyšší délka čtení, které graf dosahuje, je v řádech statisíců párů bází. Naopak nejnižší hodnota délky čtení odpovídá jednomu páru bází.

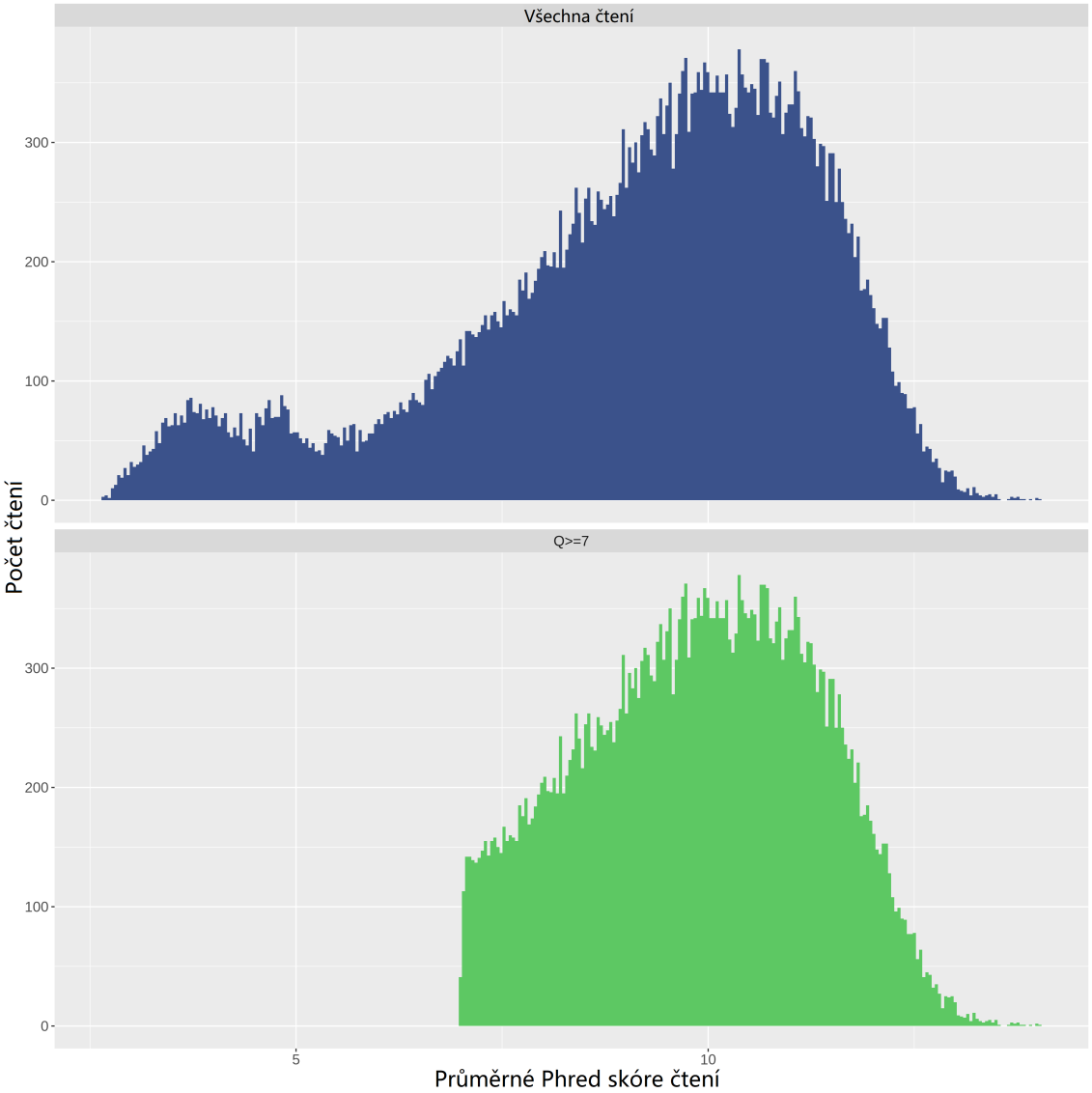
Poslední graf 5.8 vyjadřuje závislost veličiny x, jež představuje délku čtení v logaritmickém měřítku, na veličině y neboli průměrné kvalitě těchto čtení. Největší shluk bodů graf vykazuje pro čtení dlouhá přibližně 10 000 bp s kvalitou vyšší než Q = 10. Graf dále obsahuje barevné rozlišení, které vyjadřuje průměrný počet událostí (z angl. eventů) jedné báze v logaritmickém měřítku. Čím tmavší zabarvení čtení vykazuje, tím menší počet událostí na bázi obsahuje. Vlivem nízkého počtu čtení není z grafu barevné rozlišení jasně patrné.

Výstup programu MinIONQC obsahuje další grafy, které jsou obsaženy v pří-

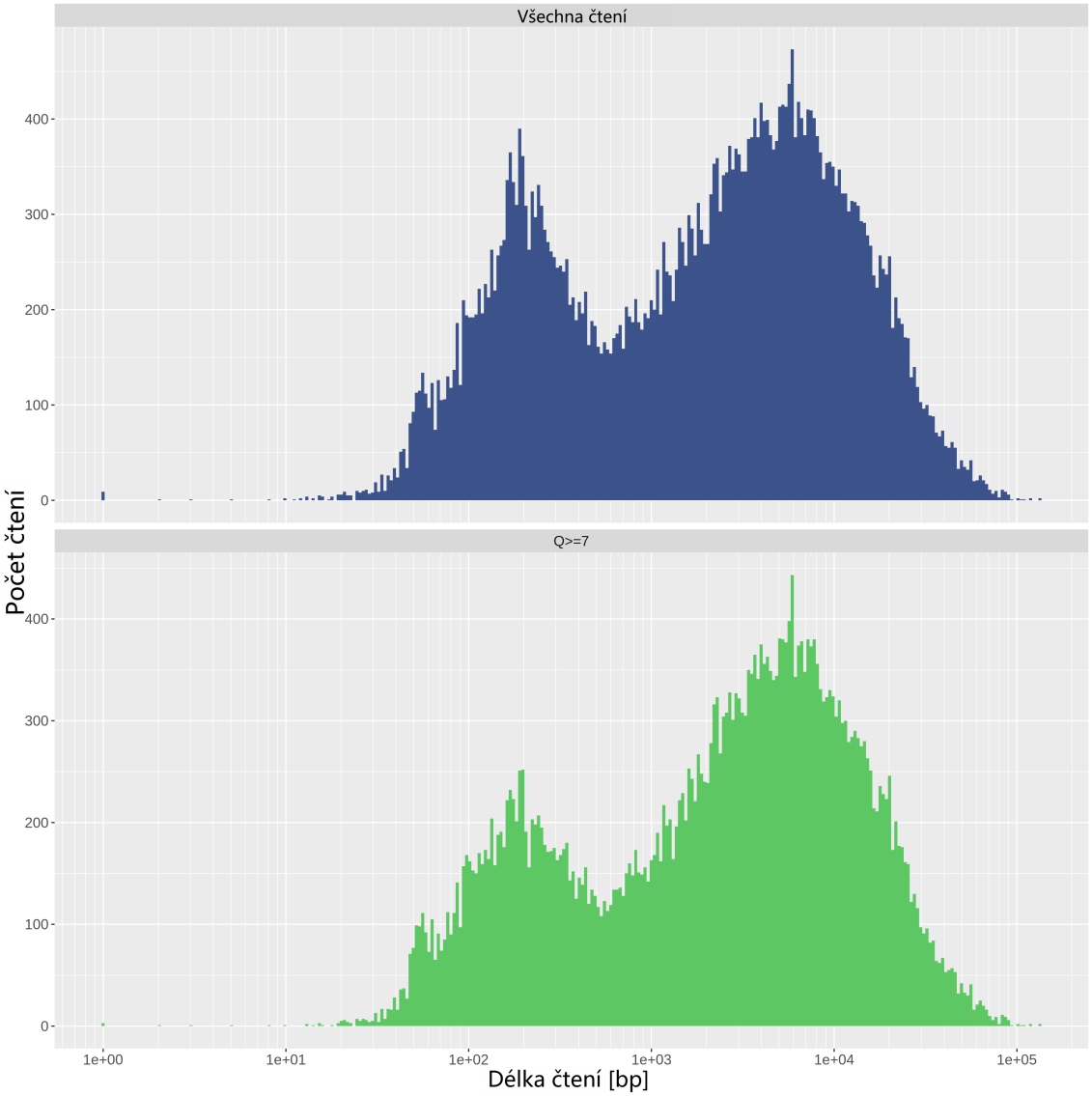
loze B. Jedná se o graf udávající počet gigabází osekvenovaných v každém kanálu sekvenační komůrky (Obr. B.1), který koresponduje s Obr. 5.5. Dále grafy vyjadřující průměr a medián délky čtení a počet bází a čtení v rámci kanálu (Obr. B.2), průměrnou délku čtení za jeden sekvenační běh (Obr. B.3), u které až do přibližně 5 hodin od začátku sekvenace docházelo k jejímu růstu až dosáhla vrcholu kolem průměrně 7 000 bp a poté začala klesat, a průměrné skóre kvality za jeden sekvenační běh (Obr. B.4), kde lze vidět, že Phred skóre začalo na vysokých hodnotách a v průběhu sekvenace došlo k jeho postupnému poklesu. Dále pak graf počtu čtení za jeden sekvenační běh (Obr. B.5), který ukazuje produkci velkého počtu čtení především na počátku sekvenace, poté grafy závislosti celkové produkce gigabází na minimální délce čtení (Obr. B.6) a celkové produkce gigabází za jeden sekvenační běh (Obr. B.7).



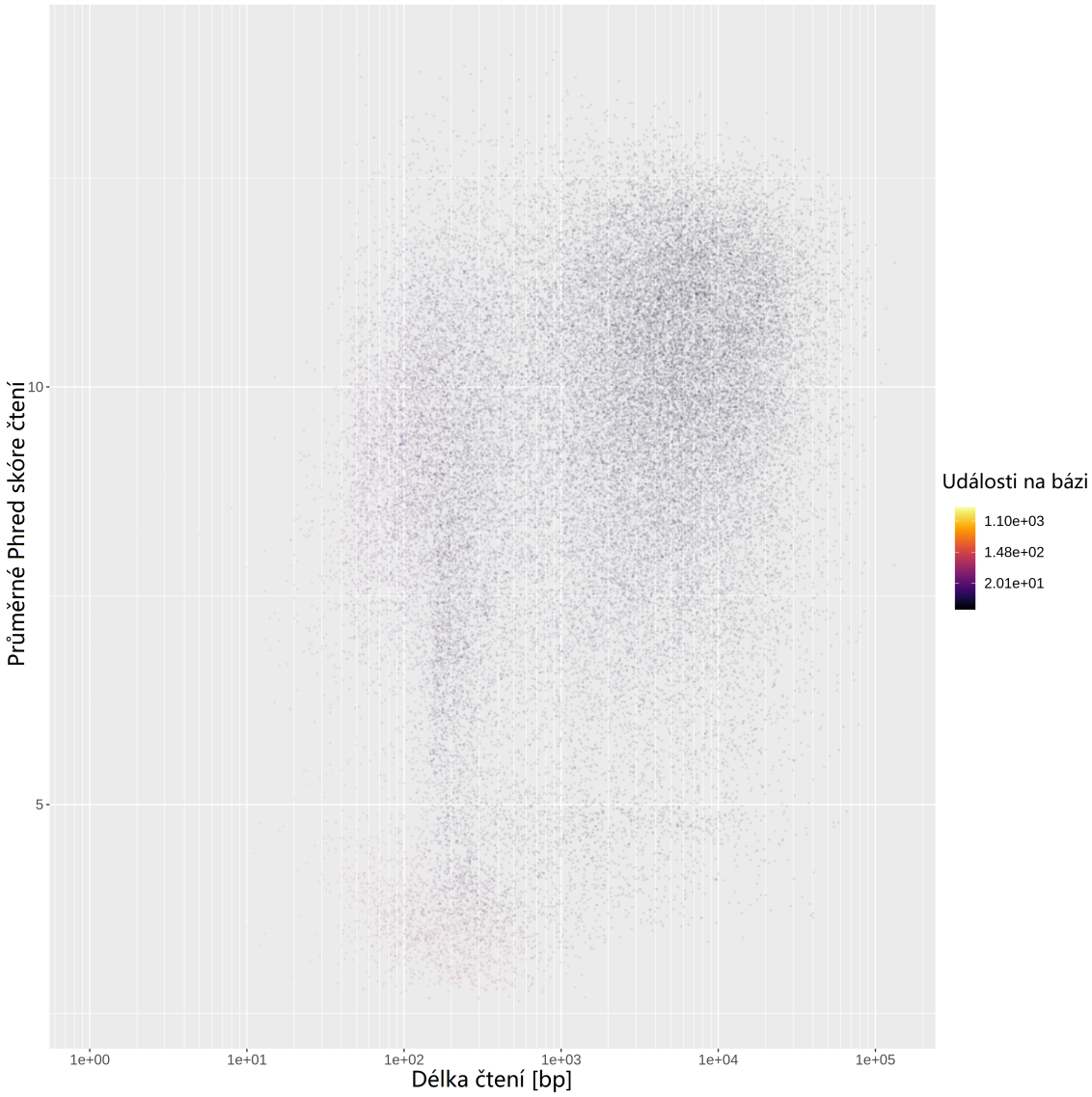
Obr. 5.5: Mapa sekvenační komůrky sekvenátoru MinION s 512 kanály pro paralelní sekvenaci. Každý kanál obsahuje dílčí graf, u něhož osa y představuje délku čtení v logaritmickém měřítku a osa x počet hodin v jednom sekvenačním běhu. Každý bod čtení navíc obsahuje informaci o jeho kvalitě v podobě zbarvení.



Obr. 5.6: Distribuce průměrné hodnoty Phred skóre vůči počtu čtení sekvenace dat pocházejících ze sekvenátoru MinION. Vrchní graf obsahuje údaje počtu čtení pro všechny hodnoty Phred skóre, zatímco spodní graf zobrazuje informace o čteních s minimální hodnotou Q = 7.



Obr. 5.7: Distribuce délky čtení v logaritmickém měřítku v rámci počtu čtení sekvenace dat pocházejících ze sekvenátoru MinION. Vrchní graf obsahuje údaje distribuce pro všechny hodnoty Phred skóre, zatímco spodní graf zobrazuje informace pouze o čteních s minimální hodnotou Q = 7.



Obr. 5.8: Poměr délky čtení v logaritmickém měřítku vůči průměrné kvalitě čtení sekvenovaných na platformě MinION. Každý bod čtení navíc obsahuje informaci o průměrném počtu událostí na bázi v podobě zbarvení.

5.2 Hodnocení kvality sestavení genomů

Po sestavení kompletních sekvencí genomů byla ohodnocena jejich kvalita. U sekvencí pocházejících z platformy Miseq byly hodnoceny soubory zarovnání ve formátu BAM, zatímco u dat z MinION byla analyzovány výsledné sekvence ve formátu FASTA.

5.2.1 Hodnocení kvality sestavení dat z Illumina Miseq

Pro hodnocení kvality sestavení dat z Illumina Miseq byl použit program QualiMap (verze 11-12-16, [30]). Do programu vstupovala zarovnaná data ve formátu BAM a výstup představuje textový soubor obsahující statistické parametry. Tyto výsledné parametry jsou uvedeny v tabulce 5.1. Tabulka obsahuje informace o počtu původních čtení, dále o počtu čtení, které prošly filtrací a byly úspěšně namapovány v rámci sestavení sekvence. Následuje informace o průměrné délce těchto čtení v jednotkách páru bází a poslední parametr činí hodnoty průměrného pokrytí, ze kterého vznikly výsledné sekvence.

Tabulka 5.1 potvrzuje tvrzení teoretické části z kapitoly 1, kde bylo zmíněno, že platforma Illumina Miseq produkuje velké množství krátkých čtení. Jak uvádí tabulka, počet čtení většiny analyzovaných genomů přesahuje 2 milióny. Tento počet se snižuje po zpracování a filtraci čtení, ke kterým došlo v rámci sestavení kompletních sekvencí, a tak je počet výsledných namapovaných čtení nižší, i když stále pro většinu genomů dosahuje hranice 2 miliónů. Průměrná délka čtení všech analyzovaných genomů se pohybuje v úzkém rozmezí od 244 bp do 249 bp. Vyšší počet čtení souvisí s posledním parametrem, jímž je průměrné pokrytí. Průměrné pokrytí se u jednotlivých genomů liší. Genom EB359 obsahuje nejnižší průměrné pokrytí, které činí 28 čtení, naopak genom EB360 obsahuje nejvyšší průměrné pokrytí rovné 196 čtením. Kromě zmíněného genomu s nejnižším průměrným pokrytím dosahují všechny genomy vysokých hodnot pokrytí a lze tedy usuzovat, že jejich sekvenace pomocí Illumina Miseq a následné sestavení výsledné sekvence DNA genomu proběhly kvalitně.

Tab. 5.1: Tabulka parametrů sestavených sekvencí z Illumina Miseq.

ID genu	Počet původních čtení	Počet namapovaných čtení	Průměrná délka čtení [bp]	Průměrné pokrytí
EB359	769 170	623 612	249	28
KP268	2 659 314	2 236 816	246	99
EB360	5 474 258	4 514 161	244	196
KP1278	2 861 674	2 136 652	249	97
KP1174	2 290 354	1 560 813	249	71
KP1268	2 961 438	2 036 805	249	92

5.2.2 Hodnocení kvality sestavení dat z ONT MinION

U dat z platformy MinION byly vyhodnoceny parametry v rámci sestavení sekvencí pomocí programu Flye, který poskytuje informace o počtu kontigů, do kterých byla čtení poskládána, dále o délce nejdelšího kontigu, parametr N50, který vyjadřuje kvalitu sestavení genomu s ohledem na spojitost a na závěr hodnotu průměrného pokrytí výslednými kontigy. Následně byla kvalita sestavení otestována pomocí zarovnání výsledných sekvencí genomů vůči referenční sekvenci NC_012731.1. K tomuto zarovnání došlo v programu MUMmer (verze 4.0.0beta2, [31]) a kvalita zarovnání byla zaznamenána jako parametr nazvaný procento zarovnání k referenci. Program MUMmer poskytl i informaci o délce výsledné sestavené sekvence. Všechny zmíněné parametry byly zaneseny do tabulky 5.2, která dále obsahuje hodnoty počtu původních čtení, jež byly získány z FASTQ souborů získaných po zpracování obdržených dat z platformy MinION.

Z tabulky 5.2 je patrné, že sekvenátor MinION produkuje nižší počet čtení než Miseq (v řádech tisíců). V rámci sestavení jsou tato čtení pospojována do větší celků, kontigů. Sníží se tak jejich počet na desítky až jednotky úseků, o čemž vypovídá parametr Počet kontigů. Délka nejdelšího kontigu se u zpracovaných dat pohybuje v rozmezí od 67 822 bp do 5 376 632 bp. S délkou kontigů souvisí parametr N50, který vyjadřuje délku kontigu, který obsahuje minimálně polovinu bází v celém sestavení. U genomů s menším počtem kontigů (genomy KP1278, KP1174, KP1268) se N50 rovná délce nejdelšího kontigu, zatímco zbylé genomy mají hodnotu N50 nižší než je délka jejich nejdelšího kontigu. Tabulka dále uvádí hodnotu průměrného pokrytí v sestavení genomu. Tato hodnota je výrazně nižší než u dat z platformy Miseq, což vyplývá z odlišných postupů sestavování genomů. Nejnižší pokrytí vykazují genomy KP268, EB360 a KP1268, jejichž průměrné pokrytí odpovídá hodnotě 3 kontigy. Nejvyšší průměrné pokrytí je rovno hodnotě 9 kontigů a náleží genomu KP1278. Délka

sestavených sekvencí se alespoň z poloviny blíží délce referenční sekvence, která činí 5 248 520 bp. Délka sekvencí KP1278 a KP1174 dokonce přesahuje délku reference přibližně o 200 000 bp. Nejkratší sekvence dosahují 130 911 bp a 853 320 bp a patří genům KP1268 a KP268. Rozdíly v délkách sekvencí se projevují i v posledním parametru, kterým je procento zarovnání sestavených sekvencí k referenční sekvenci. Nejúspěšnější je genom KP1278 s podobností referenci přes 90 %, který s ohledem na nejvyšší pokrytí a délku lze prohlásit za nejkvalitněji osekvenovaný a sestavený. Dalšími kvalitními genomy jsou KP1174 a EB359, zatímco zbylé genomy nedosahují podobnosti s referenční sekvencí ani 50 %. Nejhůře osekvenovaný a sestavený genom se nachází na posledním řádku tabulky a má označení KP1268. Jeho procento zarovnání k referenci je rovno 2,46 %, což se odvíjí od nízkého pokrytí a výsledné délky sekvence, která nedosahuje ani 0,5 miliónu párů bází.

Tab. 5.2: Tabulka základních parametrů sestavených sekvencí z ONT MinION.

ID genu	Počet původních čtení	Počet kontigů	Délka nejdelšího kontigu [bp]	N50 [bp]	Průměrné pokrytí	Délka sestavené sekvence	Procento zarovnání k referenci
EB359	2365	20	937 460	324 485	4	4 895 130	77.49
KP268	2012	13	118 345	79 454	3	853 320	13.58
EB360	3020	25	263 835	134 552	3	2 517 997	46.48
KP1278	7483	4	5 376 632	5 376 632	9	5 495 488	91.27
KP1174	6826	12	3 561 151	3 561 151	7	5 463 852	89.84
KP1268	4158	3	67 822	67 822	3	130 911	2.46

6 Metody pro porovnání bakteriálních genomů

K porovnání sestavených genomů lze použít mnoho metod komparativní genomiky. Tyto metody se v posledních letech rozvíjí a zdokonalují společně s neustálým vývojem sekvenačních technologií. Základem srovnávání genomů je porovnání jejich statistických parametrů jako je například celková délka těchto sekvencí nebo počet jednotlivých bází. Další možností jsou metody založené na zarovnání sekvencí a metody z oblasti fylogenetiky. [32]

6.1 Metody založené na zarovnání sekvencí

Zarovnání sekvencí představuje proceduru porovnávání dvou nebo více sekvencí vyhledáváním sérií znaků přítomných v sekvencích ve stejném pořadí. Zarovnání je možné provádět v globálním či lokálním měřítku. Globální je takové zarovnání, ve kterém je v každé jeho části brán ohled na celou sekvenci, zatímco lokální představuje takové zarovnání, ve kterém jsou hledány pouze nejpodobnější úseky mezi dvěma sekvencemi. Zárovnání se provádí pouze u sekvencí, u nichž je předpokládána značná podobnost.

K metodám souvisejícím se zarovnáním patří analýza bodových matic, která se používá před samotným zarovnáním, aby se rozhodlo, jaký typ bude vhodnější. Bodová matice představuje graf, na jehož osách se nachází porovnávané sekvence. V případě, že se báze sekvencí shodují, je v grafu tato shoda vyznačena tečkou. Oblasti se shodující se posloupností bází jsou tak zaznamenány jako diagonální čáry. Tímto způsobem lze porovnávat sekvence nukleotidů i proteinů.

Další metodou řadící se k zarovnání, která ale samotné zarovnání nepředstavuje, jsou substituční matice. Tato matice kvantifikuje rozdíly mezi sekvencemi a využívá se k hodnocení a nastavení skórovacího systému zarovnání.

Dalším typem využití zarovnání je základní nástroj pro vyhledávání lokálních zarovnání (z angl. Basic Local Alignment Search Tool neboli BLAST). Program BLAST porovnává nukleotidové a proteinové sekvence a vypočítává statistický význam jejich shod tak, že dotazovanou sekvenci rozdělí na menší úseky, které vyhledává v sekvencích z databáze. Podobnost úseků je hodnocena pomocí substituční matice, která poté vytváří celkové skóre zarovnání. [33]

6.2 Metody založené na fylogenetice

Fylogenetika představuje celý vědní obor, který hledá vývojové podobnosti mezi organismy a zabývá se tak studiem jejich příbuzenských vztahů. Fylogenetický vývoj bývá popsán fylogenetickými stromy. Fylogenetické metody jsou rozdělovány na metody znakové a distanční.

Znakové metody prochází každou pozici v sekvenci a určují pravděpodobnost výskytu každé báze. Mezi ně patří i vyhledání „nejlepšího možného“ stromu. Tato metoda prohledává všechny možné konstrukce stromu a vyhledá ten s největší evoluční pravděpodobností. [34]

Distanční metody vytváří distanční matice z proporcionálních vzdáleností všech dvojic sekvencí. Proporcionální vzdáleností se rozumí počet rozdílů mezi sekvencemi vůči jejich délce. Jednou z distančních metod je například metoda neváženého páru s aritmetickým průměrem (z angl. Unweighted Pair Group Method with Arithmetic mean neboli UPGMA). Jedná se o jednoduchou shlukovací metodu, která vytváří dendrogramy z matice vzdáleností. [35]

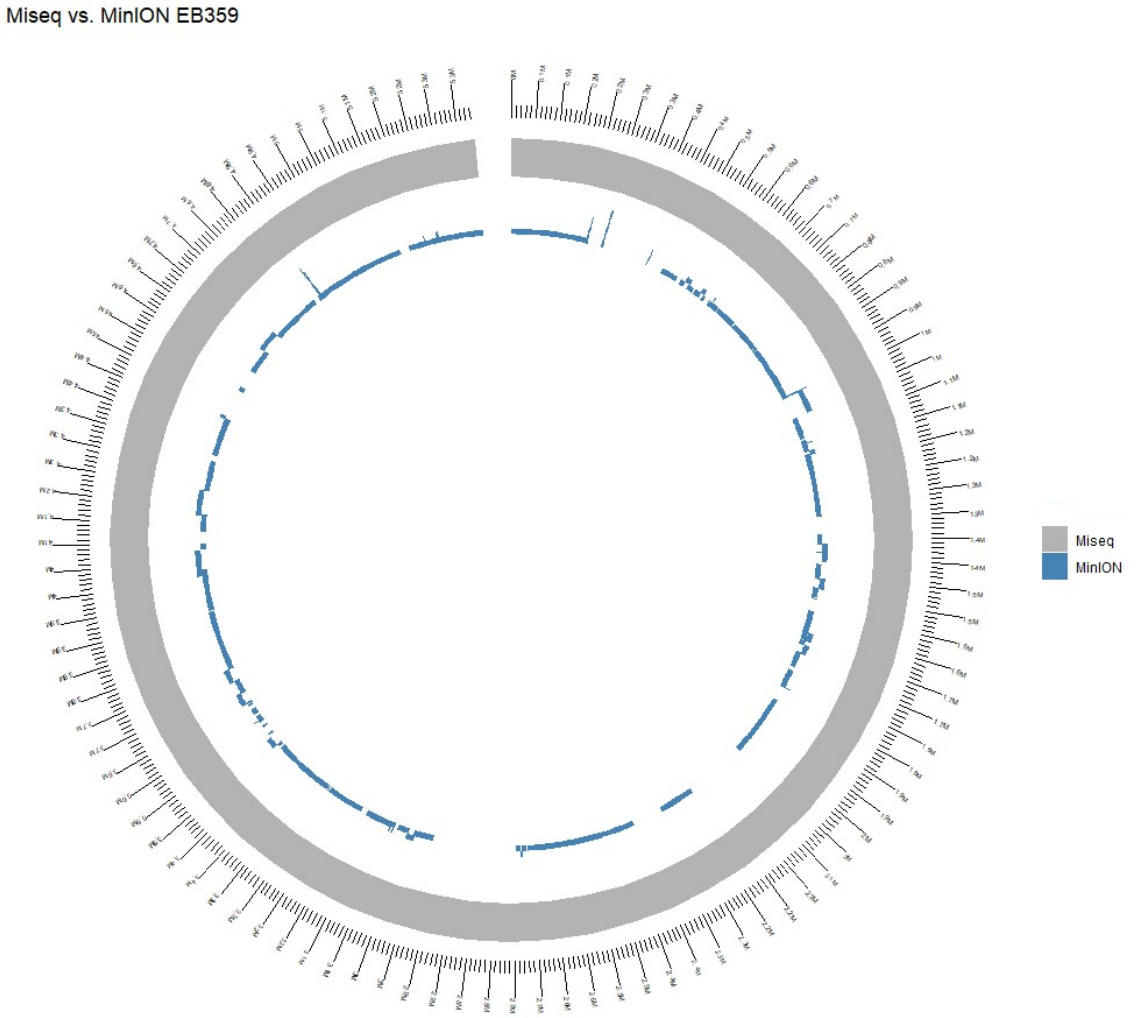
7 Porovnání sestavených bakteriálních genomů

Po sestavení bakteriálních genomů z FN Brno a ohodnocení kvality tohoto sestavení a jejich předešlé sekvenace, byly tyto genomy mezi sebou porovnávány. Nejprve došlo k zarovnání stejných sekvencí sekvenovaných na dvou odlišných sekvenačních platformách, Illumina Miseq a ONT MinION. Poté byly v genomech vyhledány geny obsažené v referenční sekvenci a na závěr byla provedena analýza kvality těchto genů v sestavených sekvencích.

7.1 Vzájemné zarovnání sestavených sekvencí

V první řadě byly vůči sobě zarovnány tytéž genomy sekvenované na dvou různých platformách, Illumina Miseq a ONT MinION. Tohoto zarovnání bylo dosaženo v programu MUMmer (verze 4.0.0beta2, [31]) a ke grafickému znázornění bylo využito programovacího prostředí a jazyka R (verze 4.0.3, [36]). Příklad grafického výstupu tohoto znázornění lze vidět na Obr. 7.1. Soubor zbylých zarovnání všech genomů je dostupný v příloze C od Obr. C.1 po Obr. C.5.

Zarovnání na obrázcích koresponduje s délkami, kterých sestavené genomy dosáhly. Všechny sekvence sekvenované pomocí platformy Miseq, které jsou na obrázcích reprezentovány šedou barvou, se shodují svojí délkou s referenční sekvencí, která má 5 248 520 bp. Této délky dosáhly díky zarovnání vůči referenci při sestavování. Přestože by podoba sekvencí v grafech mohla vyvolat dojem, že neobsahují prázdné úseky, nelze s jistotou jejich absenci předpokládat, např. ve formě nejednoznačných nukleotidů a mezer. Naopak genomy sekvenované na MinION sestavované tzv. *de novo*, které jsou vyznačeny modrou barvou, se svojí délkou liší, což lze na těchto obrázcích spatřit. Například u genomů EB360, KP268 a KP1268 na Obr. C.1, Obr. C.2 a Obr. C.4 je patrné, že u sekvencí pocházejících ze sekvenátoru MinION došlo v procesu jejich sekvenování a sestavování k velkým ztrátám úseků sekvence. Naopak Obr. 7.1, Obr. C.3 a Obr. C.5 zobrazují úspěšné zarovnání dvou sekvencí s velkým množstvím jejich překryvů.



Obr. 7.1: Grafické znázornění zarovnání sekvence EB359 sekvenované pomocí platformy Miseq a sekvence EB359 sekvenované na platformě MinION.

7.2 Vyhledávání genů v sestavených sekvencích

Jako další krok porovnávání sestavených sekvencí byly v genomech vyhledány geny, které jsou obsaženy v referenční sekvenci *Klebsiella pneumoniae* NC_012731.1 získané z databáze RefSeq NCBI, pomocí BLAST+ (verze 2.10.0, [37]). Tento program slouží pro vyhledávání homologních sekvencí a to tak, že každý gen zarovnal vůči sekvenci genomu s určitým skóre zarovnání a e-hodnotou, která vyjadřuje předpokládaný počet nalezených sekvencí o stejné nebo lepší podobnosti v databázi náhodných sekvencí o stejné velikosti, jako je dotazovaná reálná databáze. Čím menší je tedy e-hodnota, tím důvěryhodnější je výsledek zarovnání. Tyto parametry jsou znázorněny v tabulce 7.1.

Tato tabulka také udává, kolik genů bylo pomocí BLAST nalezeno v jednotlivých sekvencích. Celkový počet genů v referenční sekvenci je 5 282. Tento počet se nepodařilo najít ani v jedné ze sestavených sekvencí, avšak nejvíce se tomuto číslu přiblížil genom KP1268 sekvenovaný pomocí platformy Illumina Miseq, který obsahuje 4 922 genů. U všech genomů sekvenovaných pomocí tohoto sekvenátoru bylo nalezeno přes 4 800 genů, zatímco u genomů sekvenovaných pomocí MinION se počet genů výrazně lišil. Nejvyššího počtu za genomy sekvenované pomocí této platformy dosáhla sekvence KP1278, která obsahuje 4 780 genů. Naopak nejnižšího počtu dosáhl genom KP1268 sekvenovaný platformou MinION, a to pouze 139 genů. Dalším genomem s nízkým počtem genů je KP268, který obsahuje 703 genů. Především tyto genomy obsahují velký nepoměr v počtu nalezených genů mezi sekvencemi sekvenovaných platformou Miseq a MinION. Dále sekvence genomu EB360 získaná z MinION obsahuje přibližně polovinu počtu genů, které jsou obsaženy ve stejném genomu sekvenovaném pomocí platformy Miseq.

Další sloupec této tabulky je zaměřen na již vysvětlenou e-hodnotu. Je uveden počet genů nalezených v sekvenci, které vykazovaly e-hodnotu rovnou nule a jejichž výsledky byly tedy nejdůvěryhodnější. Opět takových genů lze nalézt více v sekvencích získaných pomocí platformy Illumina Miseq, zatímco například genomy KP268 a KP1268 sekvenované na MinION obsahují pouze 524 a 98 genů z celkových 5 282, jejichž vyhledávání nenalezlo žádnou další odpovídající sekvenci v dané databázi.

Posledním parametrem v tabulce 7.1 je již zmíněné skóre, které vyjadřuje míru podobnosti nukleotidů nalezeného genu v sestavené sekvenci s genem z referenční sekvence. Čím je tedy skóre vyšší, tím více se geny podobají. Tento parametr charakterizuje průměrné skóre vypočítané jako průměr skóre všech genů v genomu. Nejvyššího průměrného skóre dosahuje sekvence z platformy Miseq genomu EB359, zatímco geny s nejnižším průměrným skóre zarovnání obsahuje sekvence KP1268 pochazející z MinION. V této sekvenci se tedy nachází pouhých 139 genů, které jsou navíc nejméně podobné referenčním genům. Pokud jsou brány v úvahu zvlášť geny ze

všech sekvencí získaných z různých platforem, všechna průměrná skóre genů sekvencí ze sekvenátoru MinION jsou menší než průměrná skóre genů sekvencí z platformy Miseq.

Tab. 7.1: Tabulka shrnující informace z výstupu vyhledávání genů v jednotlivých genomech pomocí BLAST.

ID genomu	Sekvenátor	Počet nalezených genů v sekvenci	Počet genů s e-hodnotou = 0	Průměrné skóre
EB359	Miseq	4 823	4 011	1 620
	MinION	4 031	3 212	1 404
EB360	Miseq	4 910	4 060	1 604
	MinION	2 471	1 910	1 319
KP268	Miseq	4 852	4 021	1 615
	MinION	703	524	1 254
KP1174	Miseq	4 883	4 041	1 606
	MinION	4 672	3 809	1 518
KP1268	Miseq	4 922	4 005	1 586
	MinION	139	98	1 249
KP1278	Miseq	4 855	3 984	1 610
	MinION	4 780	3 912	1 542

7.3 Porovnání sestavených sekvencí na základě nalezených genů

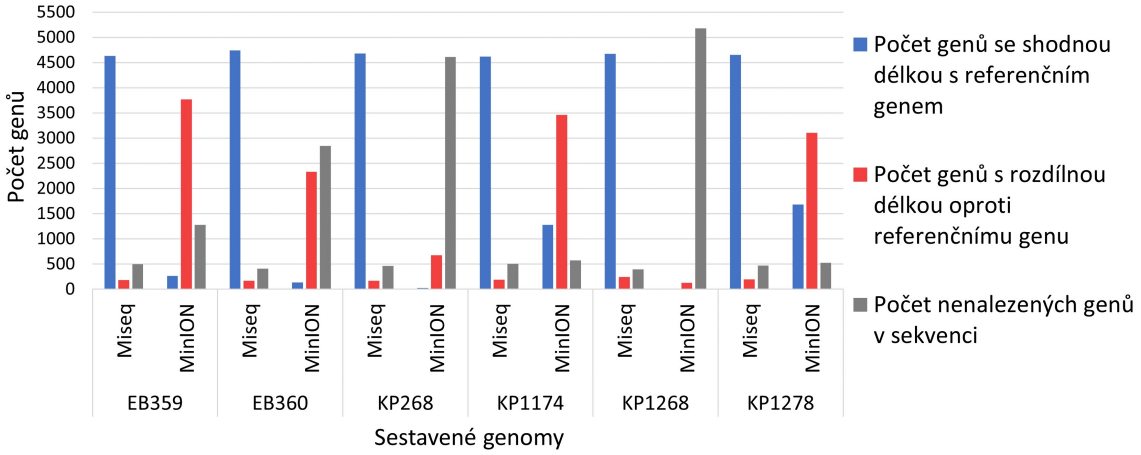
Po nalezení genů z referenční sekvence v sestavených genomech byl v programovacím protředí MATLAB (verze R2020a, [38]) vytvořen kód pro porovnání sekvencí stejného genomu sekvenovaných na různých platformách z hlediska délek genů v nich nalezených a dále z hlediska bodových mutací jako jsou substituce, inzerce a delece. Byly použity výstupy z vyhledávání pomocí BLAST, ze kterých byly vyjmuty nalezené úseky posloupnosti nukleotidů.

Jak bylo zmíněno, v prvé řadě byly porovnávány délky nalezených genů. Vytvořený kód, jehož schéma lze vidět v příloze D na Obr. D.1, zjistil délku nalezeného úseku a porovnal ji s délkou odpovídajícího genu z referenční sekvence. Výsledky tohoto porovnání jsou znázorněny na Obr. 7.2 a poté detailněji pro každý genom zvlášť v grafech v příloze E od Obr. E.1 až po Obr. E.6. Grafy pro názornější srovnání obsahují také informaci o počtu genů, které nebyly v sestaveném genomu nalezeny. Grafy byly vytvořeny z hodnot v tabulce F.1, která je umístěna v příloze F.

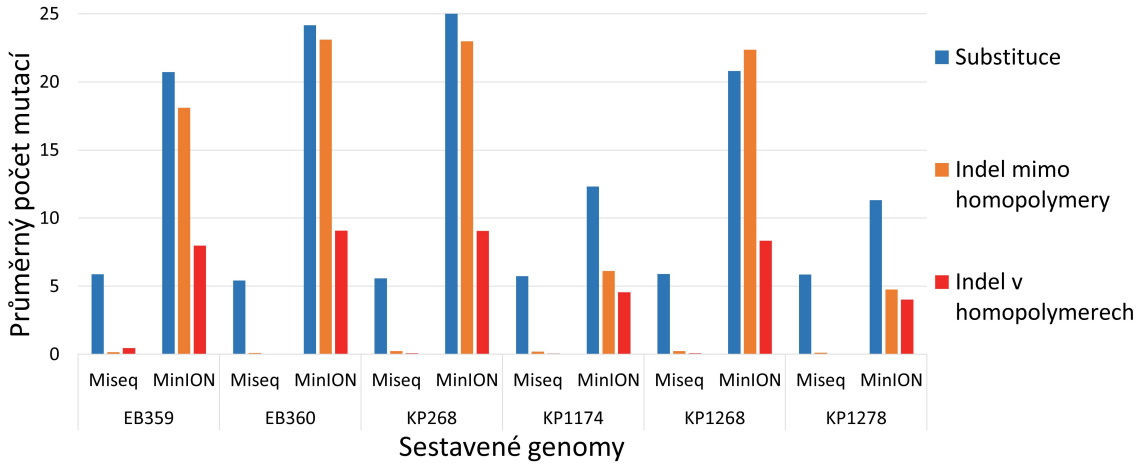
Obr. 7.2 znázorňuje skupinový sloupcový graf vyjadřující porovnání počtu genů ve všech genomech sekvenovaných na platformách Miseq a MinION. Z grafu je patrné, že sekvence získané pomocí platformy Miseq obsahují velký počet genů o stejné délce jako referenční sekvence, zatímco geny v sekvencích genomu sekvenovaného pomocí MinION se svojí délkou do velké míry liší. Poslední sloupec šedé barvy také udává, že v genomech sekvenovaných na MinION nebyl nalezen větší počet genů než v genomech získaných pomocí Miseq. Tato informace je patrná především pro sekvence EB360, KP268 a KP1268 na Obr. E.2, Obr. E.3 a Obr. E.5, kde počet nenalezených genů v sekvenci z MinION výrazně převyšuje počet genů nalezených. Všechny grafy vykazují velmi podobné rozložení sloupců pro všechny genomy sekvenované na sekvenátoru Miseq. Modrý sloupec počtu genů se shodnou délkou s referenční sekvencí u všech těchto genomů přesahuje hranici 4 500 genů, zatímco červený sloupec s rozdílnými délkami v žádném z případů nedosahuje ani 500 genů. Podobně jako červený sloupec se chová i šedý sloupec s nenalezenými geny. Naopak části grafů patřící platformě MinION se velmi liší. Avšak všechny grafy obsahují červený sloupec značící geny s rozdílnou délkou převyšující modrý sloupec se shodujícími se délkami. Tento rozdíl by mohl značit větší výskyt inzercí a delecí v sekvencích pocházejících ze sekvenátoru ONT MinION. Nejvyšších hodnot modrého sloupce značícího počet genů se shodnou délkou jako gen z referenční sekvence z genomů pocházejících ze sekvenátoru MinION dosahují sekvence KP1174 a KP1278 na Obr. E.4 a Obr E.6, a to přes 1 000 a 1 500 genů, zároveň však spolu s genomem EB359 na Obr E.1 přesahují hranici 3 000 genů, které se naopak svojí délkou od reference liší.

Další porovnání bylo zaměřeno na výskyt substitucí, inzercí a delecí. K tomu byl využit algoritmus, který prošel zarovnání genů nalezených ve stejných genomech z různých sekvenátorů, vyhledal pozice, ve kterých se liší a dle podoby nukleotidů nacházejících se na těchto pozicích určil, o jaký druh bodové mutace se jedná. V případě inzercí a delecí kód dále zjišťoval, zda se vyskytují samostatně nebo jsou součástí homopolymerů, tedy úseků sekvence, které obsahují pouze nukleotidy stejného typu. Toho bylo dosaženo tak, že kód porovnal předpokládaný nukleotid na pozici mutace z referenční sekvence s nukleotidy na pozici obklopující mutaci v sestavené sekvenci. Schéma algoritmu je zobrazeno na v příloze D na Obr. D.2. Výsledky tohoto porovnání jsou zobrazeny na Obr. 7.3 a dále detailněji pro každý genom zvlášť v grafech v příloze E od Obr. E.7 až po Obr. E.12. Grafy byly vytvořeny z hodnot v tabulce F.2, která se nachází v příloze F.

Obrázky představují sloupcové grafy pro stejné genomy sekvenované pomocí dvou platforem. Pro každý takový genom jsou vyobrazeny tři sloupce. První modrý sloupec vyjadřuje průměrný počet substitucí v nalezených genech, druhý oranžový sloupec udává průměrný počet inzercí či delecí v genu, které se nenacházely v rámci homopolymerů, a třetí červený sloupec obsahuje průměrný počet inzercí a delecí, které naopak vznikly jako součást homopolymerů. Každý z grafů vykazuje vyšší sloupce v části, která odpovídá genomům získaným z platformy MinION. Modrý sloupec u všech takových genomů kromě KP1174 a KP1278 (Obr. E.10 a Obr. E.12) přesahuje průměrnou hodnotu 20 substitucí v jednom genu. Dva zbývající zmíněné genomy přesahují průměrný počet 10 substitucí. Pro sekvence z Miseq je modrý sloupec vždy nejvyšší, avšak v žádném z případů nedosahuje ani 6 průměrných substitucí v rámci jednoho genu. Co se týče dalších dvou sloupců patřících inzercím a delecím, žádný z genomů této platformy nedosahuje ani hodnoty 1, což odpovídá výsledkům měření délek genů, které udávaly velký počet shodných délek s geny z referenční sekvence. Na druhou stranu inzerce a delece hrají velkou roli u sekvencí z MinION. Opět kromě sekvencí KP1174 a KP1278, jejichž indely se v homopolymerech i mimo ně pohybují kolem průměrného počtu 5 inzercí/delecí na jeden gen, přesahují zbývající sekvence z MinION, tedy EB359, EB360, KP268 a KP1268 (na Obr. E.7, Obr. E.8, Obr. E.9 a Obr. E.11), hranici průměrného počtu 15-20 inzercí/delecí mimo homopolymery. Co se týče těchto sekvencí a indelů v rámci homopolymerů obsažených v jejich genech, pohybují se mezi průměrnou hodnotou 8-9 těchto bodových mutací na jeden gen. Vezme-li se v úvahu počet nalezených genů v sekvencích KP1174 a KP1278 a nízký průměrný počet bodových mutací v těchto genech, jsou tyto genomy svojí kvalitou nejblíže sekvencím pocházejícím ze sekvenátoru Illumina Miseq.



Obr. 7.2: Graf porovnání počtu nalezených genů v rámci všech genomů z hlediska jejich délek.



Obr. 7.3: Graf porovnání průměrného počtu bodových mutací v nalezených genech v rámci všech genomů.

8 Diskuze a vyhodnocení porovnávání sestavených genomů

K tomu, aby bylo možné porovnat bakteriální genomy sekvenované pomocí druhé a třetí generace sekvenátorů, bylo zapotřebí provést několik kroků. Nejprve bylo potřeba tyto genomy sestavit a zhodnotit kvalitu tohoto sestavení a předcházející sekvenace. V rámci samotného porovnání došlo k zarovnání stejných genomů sekvenovaných pomocí jiných platforem vůči sobě, následně byly ve všech genomech vyhledány geny, které jsou obsaženy v referenční sekvenci, a na závěr byla provedena analýza nalezených genů z hlediska jejich délek a bodových mutací v nich obsažených.

Na základě těchto kroků je možné vyhodnotit porovnání bakteriálních genomů pocházejících z různých generací sekvenátorů. Z hodnocení sekvenací bylo zjištěno, že oba tyto procesy proběhly poměrně kvalitně, avšak u sekvenátoru MinION nebyla využita velká část kanálů sekvenační komůrky, což mohl být důvod pro nižší hodnoty Phred skóre čtení produkovaných touto platformou stejně jako pro nižší počet čtení, které tato platforma produkovala. Při hodnocení po sestavení genomů bylo odhaleno, že pouze tři z šesti genomů sekvenovaných na platformě MinION se svojí délkou blíží referenční sekvenci narozdíl od genomů pocházejících ze sekvenátoru Miseq, které bez výjimky dané délky dosahují. Tento rozdíl mohl být z velké části zaviněn faktem, že genomy pocházející z jiných platforem byly sestaveny rozdílnými způsoby. Sekvence z platformy Miseq byly sestavovány vůči referenci, avšak sekvence z platformy MinION byly sestaveny tzv. *de novo*, tedy bez přispění referenční sekvence. Tyto odlišnosti byly graficky znázorněny také v kapitole 7.1 při zarovnávání dvou stejných genomů pocházejících z různých sekvenačních platforem vůči sobě. Při vyhledávání genů se ukázalo, že nedostatečná délka některých sekvencí je propojená s nenalezením těchto genů. Zatímco u všech genomů z platformy Miseq nebyl problém najít přes 90% genů z reference, opět jen polovina sekvencí získaných pomocí MinION se dokázala přiblížit této úspěšnosti. Fakt, že došlo k nalezení genů, nebyl předpokladem pro shodnost genu s referencí, proto byla dále rozhodující otázka kvality těchto genů a jejich podobnost s geny obsaženými v referenci. O tomto vypovídá průměrné skóre z Tab. 7.1, které opět vychází lépe pro sekvence z platformy Miseq. Při porovnávání délek genů bylo zjištěno, že tyto sekvence obsahují vyšší počet genů se shodnou délkou s geny z referenční sekvence. Z tohoto poznatku bylo možno získat předpoklad, že geny v genomech pocházejících ze sekvenátoru Miseq obsahují nižší počet inzercí a delecí. Toto zjištění bylo potvrzeno a graficky znázorněno na Obr. 7.3. Při analýze bodových mutací v nalezených genech byl objeven výrazný problém v sekvencích ze sekvenátoru MinION týkající se inzercí a delecí jak

mimo homopolymerní úseky, tak i v nich. Právě neschopnost sekvenátoru MinION zachytit přesnou délku homopolymerních sekvencí je předmětem mnohých článků z odborné literatury, které tento nedostatek připisují procesu rozkódování signálu zaznamenávaného na nanopórech, který je snímán po pěticích nukleotidů. Kromě toho tyto sekvence překonávají genomy pocházející z platformy Miseq také v průměrném počtu substitucí v každém nalezeném genu. Zatímco sekvence platformy Miseq obsahují v průměru 5 substitucí v každém genu oproti referenci, i sekvence platformy MinION, které obdržely nejvyšší skóre při zarovnání genů vůči referenci, překonávají tento průměrný počet nejméně o dalších 6 substitucí.

Na základě provedených analýz lze prohlásit, že ať již rozdíly mezi genomy sekvenovanými odlišnými platformami byly zapříčiněny chybovostí sekvenátoru MinION či odlišnými přístupy sestavování sekvencí, je výsledná podoba sekvencí získaných z platformy druhé generace, Illumina Miseq, mnohem kvalitnější a více podobná referenční sekvenci než sekvence pocházející ze sekvenátoru třetí generace, Oxford Nanopore Technologies MinION, u nichž i genomy s nejlepšími výsledky nedosahují kvality genomů platformy Miseq.

Závěr

Náplní této bakalářské práce je vypracování literární rešerše na téma sekvenačních technologií, sestavení genomů ze čtení produkovaných sekvenátory druhé a třetí generace, otestování kvality tohoto sestavení a předchozí sekvenace a návrh a implementace metody pro porovnání bakteriálních genomů.

Úvod teoretické části obsahuje popis sekvenačních technologií a platforem s podrobným zaměřením na sekvenátory Illumina Miseq a Oxford Nanopore Technologies MinION. Dále je navázáno objasněním postupu sestavení genomu. Uvedeny jsou dva algoritmy, které jsou detailně popsány. Teoretická část je uzavřena popisem struktury sekvenačních dat a formátů, ve kterých bývají zapsány.

V další části došlo k sestavení kompletních sekvencí genomů bakterie *Klebsiella pneumoniae* poskytnutých z Fakultní nemocnice Brno. Data byla osekvenována na dvou platformách, Illumina Miseq a Oxford Nanopore Technologies MinION, na které byl zaměřen začátek teoretické části. U dat pocházejících ze sekvenátoru Miseq bylo nejprve pomocí softwaru BWA provedeno referenční mapování, ze kterého byla poté užitím sady nástrojů SAMtools vygenerována konsenzuální sekvence a převedena do podoby formátu FASTA. Naopak data z platformy MinION musela nejdříve projít převedením do znakové podoby a odstřižením identifikátorů představujících umělou DNA přidanou před sekvenováním. Až poté byla čtení těchto dat sestavena tzv. *de novo* a došlo k vytvoření výsledného FASTA souboru.

Praktická část pokračuje testováním kvality poskytnutých genomů. Nejprve došlo k otestování kvality čtení sekvencí při sekvenování na platformě Miseq pomocí programu FastQC. Tento program vyhodnotil např. Phred skóre jednotlivých čtení a zaznamenal distribuci jeho hodnot na délku čtení sekvence do grafů. Výsledkem tohoto testování jsou vysoká skóre kvality sekvencí z platformy Miseq. Dále došlo k hodnocení kvality sekvenace dat z platformy MinION pomocí programu MinIONQC, který poskytnul grafy zobrazující např. fyzické rozložení kanálů sekvenační komůrky a jejich reálné využití při sekvenaci, dále pak grafy zabývající se délkou a kvalitou jednotlivých čtení. Zde bylo zjištěno nevyužití celého povrchu sekvenační komůrky při sekvenaci, což mohlo způsobit nižší hodnoty kvality čtení a počtu produkovaných čtení.

V rámci testování kvality genomů byla následně hodnocena kvalita jejich sestavení. Výsledky pro obě platformy byly zaznamenány do tabulek, které byly zaměřené především na porovnání délky a počtu čtení před sestavením a následně po něm. Nechyběla ani informace o průměrném pokrytí. Výsledná data pro dva odlišné sekvenátory se lišila především z důvodu využití odlišných postupů sestavení genomů.

Závěrečná část bakalářské práce obsahuje nejprve teoretické pojednání o metodách, které lze použít k porovnání sekvencí genomů, v němž jsou zmíněny dva

přístupy. Navazuje praktické využití těchto metod. Nejprve byly vůči sobě zarovnány stejné sekvence pocházející z odlišných platforem. Výsledek tohoto zarovnání byl zobrazen v grafech, v nichž bylo možné pozorovat chybějící úseky v sekvencích pocházejících ze sekvenátoru MinION. Dalším krokem závěrečné analýzy bylo vyhledání genů obsažených v referenční sekvenci v sestavených genomech. Tento postup byl proveden pomocí BLAST a výsledky zaznamenány do tabulky, kde bylo zjištěno, že všechny genomy sekvenované na Miseq obsahují větší počet vyhledávaných genů než sekvence získané pomocí MinION. Následovala analýza kvality těchto genů. V první řadě byla zkontrolována délka nalezených genů a na závěr bylo odhaleno, zda obsahují bodové mutace. Výsledky tohoto rozboru uvádějí vyšší podobnost genů obsažených v genomech pocházejících ze sekvenátoru Miseq s geny v referenční sekvenci.

Poslední kapitola obsahuje kompletní shrnutí všech hodnocení a analýz provedených v této bakalářské práci, z nichž je vyvozen výsledek porovnání bakteriálních genomů z druhé a třetí generace sekvenátorů společně s možnými důvody objasňujícími vzniklé rozdíly mezi sekvencemi těchto platforem. Z tohoto porovnání vychází jako kvalitnější a referenci podobnější genomy získané sekvenací na platformě Illumina Miseq.

Literatura

[1] METZKER, M. L. Emerging technologies in DNA sequencing. Genome Research. 2005, 15(12), 1767-1776. ISSN 1088-9051. Dostupné z: doi:10.1101/gr.3770505

[2] HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016, 107(1), 1-8. ISSN 08887543. Dostupné z: doi:10.1016/j.ygeno.2015.11.003

[3] GUPTA, Nidhi a Vijay K. VERMA. Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. Microbial Technology for the Welfare of Society. Singapore: Springer Singapore, 2019, 2019-09-13, , 313-341. Microorganisms for Sustainability. ISBN 978-981-13-8843-9. Dostupné z: doi:10.1007/978-981-13-8844-6_15

[4] BEHJATI, Sam a Patrick S TARPEY. What is next generation sequencing?: The history of sequencing DNA. Genomics. 2013, 98(6), 236-238. ISSN 1743-0585. Dostupné z: doi:10.1136/archdischild-2013-304340

[5] BUERMANS, H.P.J. a J.T. DEN DUNNEN. Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2014, 1842(10), 1932-1941. ISSN 09254439. Dostupné z: doi:10.1016/j.bbadis.2014.06.015

[6] METZKER, Michael L. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010, 11(1), 31-46. ISSN 1471-0056. Dostupné z: doi:10.1038/nrg2626

[7] BROWN, T., BROWN (Jr), T., Nucleic Acids Book, ATDBIO. Dostupné z: https://www.atdbio.com/nucleic-acids-book

[8] CAPORASO, J Gregory, Christian L LAUBER, William A WALTERS, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal. 2012, 6(8). ISSN 1751-7362. Dostupné z: doi:10.1038/ismej.2012.8

[9] ZVÁROVÁ, Jana a Ivan MAZURA. Metody molekulární biologie a bioinformatiky. Praha: Karolinum, 2012. Biomedicínská informatika. ISBN 978-80-246-2150-0.

[10] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and Genome Assembly. 2016, 14(5), 265-279. ISSN 16720229. Dostupné z: doi:10.1016/j.gpb.2016.05.004

[11] TYLER, Andrea D., Laura MATASEJE, Chantel J. URFANO, Lisa SCHMIDT, Kym S. ANTONATION, Michael R. MULVEY a Cindi R. CORBETT. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Scientific Reports. 2018, 8(1). ISSN 2045-2322. Dostupné z: doi:10.1038/s41598-018-29334-5

[12] PAREEK, Chandra Shekhar, Rafal SMOCZYNSKI a Andrzej TRETYN. Sequencing technologies and genome sequencing. Journal of Applied Genetics. 2011, 52(4), 413-435. ISSN 1234-1983. Dostupné z: doi:10.1007/s13353-011-0057-x

[13] COMPEAU, Phillip E C, Pavel A PEVZNER a Glenn TESLER. How to apply de Bruijn graphs to genome assembly. Nature Biotechnology. 2011, 29(11), 987-991. ISSN 1087-0156. Dostupné z: doi:10.1038/nbt.2023

[14] HENSON, Joseph, German TISCHLER a Zemin NING. Next-generation sequencing and large genome assemblies. Pharmacogenomics. 2012, 13(8), 901-915. ISSN 1462-2416. Dostupné z: doi:10.2217/pgs.12.72

[15] SCHATZ, M. C., A. L. DELCHER a S. L. SALZBERG. Assembly of large genomes using second-generation sequencing. Genome Research. 2010, 20(9), 1165-1173. ISSN 1088-9051. Dostupné z: doi:10.1101/gr.101360.109

[16] JAIN, Miten, Hugh E. OLSEN, Benedict PATEN a Mark AKESON. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology. 2016, 17(1). ISSN 1474-760X. Dostupné z: doi:10.1186/s13059-016-1103-0

[17] LI, Z., Y. CHEN, D. MU, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in Functional Genomics. 2012, 11(1), 25-37. ISSN 2041-2649. Dostupné z: doi:10.1093/bfgp/elr035

[18] COMMINS, Jennifer, Christina TOFT a Mario A. FARES. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. Biological Procedures Online. 2009, 11(1), 52-78. ISSN 1480-9222. Dostupné z: doi:10.1007/s12575-009-9004-1

[19] LOMAN, Nicholas J, Joshua QUICK a Jared T SIMPSON. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods. 2015, 12(8), 733-735. ISSN 1548-7091. Dostupné z: doi:10.1038/nmeth.3444

[20] COCK, Peter J. A., Christopher J. FIELDS, Naohisa GOTO, Michael L. HEUER a Peter M. RICE. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010, 38(6), 1767-1771. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkp1137

[21] LI, H., B. HANDSAKER, A. WYSOKER, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009, 25(16), 2078-2079. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btp352

[22] BOLGER, Anthony M., Marc LOHSE a Bjoern USADEL. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014, 30(15), 2114-2120. ISSN 1460-2059. Dostupné z: doi:10.1093/bioinformatics/btu170

[23] LI, H., R. DURBIN, Yu LIN a Pavel A. PEVZNER. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009, 25(14), 1754-1760. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btp324

[24] DANECEK, P., A. AUTON, G. ABECASIS, et al. The variant call format and VCFtools. Bioinformatics. 2011, 27(15), 2156-2158. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btr330

[25] OXFORD NANOPORE TECHNOLOGIES. Guppy Protocol. 2020. [Online]. Dostupné z: https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revt_14dec2018

[26] WICK, Ryan R., Louise M. JUDD a Kathryn E. HOLT. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology. 2019, 20(1). ISSN 1474-760X. Dostupné z: doi:10.1186/s13059-019-1727-y

[27] KOLMOGOROV, Mikhail, Jeffrey YUAN, Yu LIN a Pavel A. PEVZNER. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. 2019, 37(5), 540-546. ISSN 1087-0156. Dostupné z: doi:10.1038/s41587-019-0072-8

[28] ANDREWS, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Dostupné z: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[29] LANFEAR, R, M SCHALAMUN, D KAINER, W WANG, B SCHWESSINGER a John HANCOCK. MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics. 2019, 35(3), 523-525. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/bty654

[30] GARCÍA-ALCALDE, Fernando, Konstantin OKONECHNIKOV, José CARBONELL, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012, 28(20), 2678-2679. ISSN 1460-2059. Dostupné z: doi:10.1093/bioinformatics/bts503

[31] MARÇAIS, Guillaume, Arthur L. DELCHER, Adam M. PHILLIPPY, Rachel COSTON, Steven L. SALZBERG, Aleksey ZIMIN a Aaron E. DARLING. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018, 14(1). ISSN 1553-7358. Dostupné z: doi:10.1371/journal.pcbi.1005944

[32] CVRČKOVÁ, Fatima. Úvod do praktické bioinformatiky. Praha: Academia, 2006. ISBN 80-200-1360-1.

[33] BAXEVANIS, Andreas D. a B. F. Francis OUELLETTE. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. SECOND EDITION. Hardcover; Hoboken, New Jersey, U.s.a: Wiley-Interscience, October 29, 2004. ISBN 978-0471478782.

[34] PRIMROSE, Sandy B. a Richard M. TWYMAN. Principles of Genome Analysis and Genomics. THIRD EDITION. Blackwell Science a Blackwell Publishing company, 2003. ISBN 1-40510-120-2.

[35] STORMO, G D. Computer Methods for Analyzing Sequence Recognition of Nucleic Acids. Annual Review of Biophysics and Biophysical Chemistry. 1988, 17(1), 241-263. ISSN 0883-9182. Dostupné z: doi:10.1146/annurev.bb.17.060188.001325

[36] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[37] ALTSCHUL, Stephen F., Warren GISH, Webb MILLER, Eugene W. MYERS a David J. LIPMAN. Basic local alignment search tool. Journal of Molecular Biology. 1990, 215(3), 403-410. ISSN 00222836. Dostupné z: doi:10.1016/S0022-2836(05)80360-2

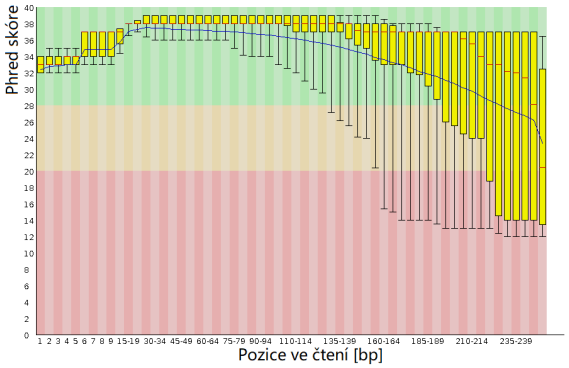
[38] MATLAB, 2020. 9.8.0.1359463 (R2020a), Natick, Massachusetts: The MathWorks Inc.

Seznam symbolů, veličin a zkratek

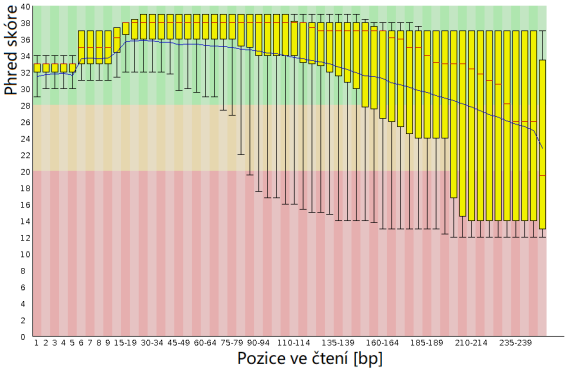
ASCII	American Standard Code for Information Interchange
ASIC	Application Specific Integrated Circuit
BLAST	Basic Local Alignment Search Tool
CIGAR	Concise Idiosyncratic Gapped Alignment Report
DNA	deoxyribonukleová kyselina
dNTP	deoxynukleosidtrifosfát
IUPAC	International Union of Pure and Applied Chemistry
NCBI	National Center for Biotechnology Information
PCR	polymerázové řetězové reakce
SMRT	Single molecule real time sequencing
UPGMA	Unweighted Pair Group Method with Arithmetic mean

Seznam příloh

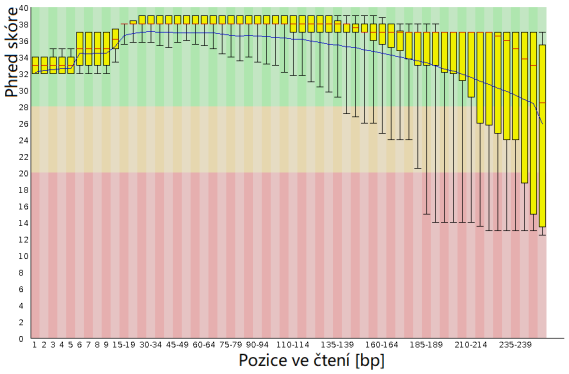
A Výstup z FastQC



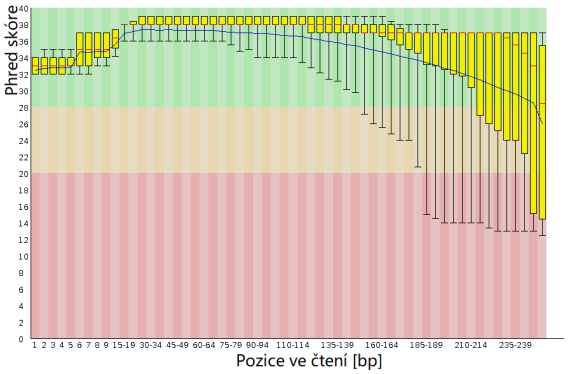
(a) Genom KP268



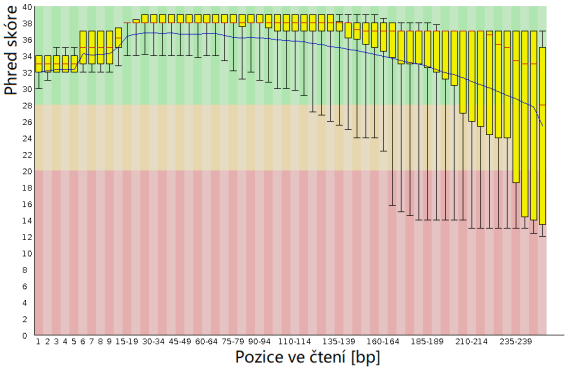
(b) Genom EB360



(c) Genom KP1278

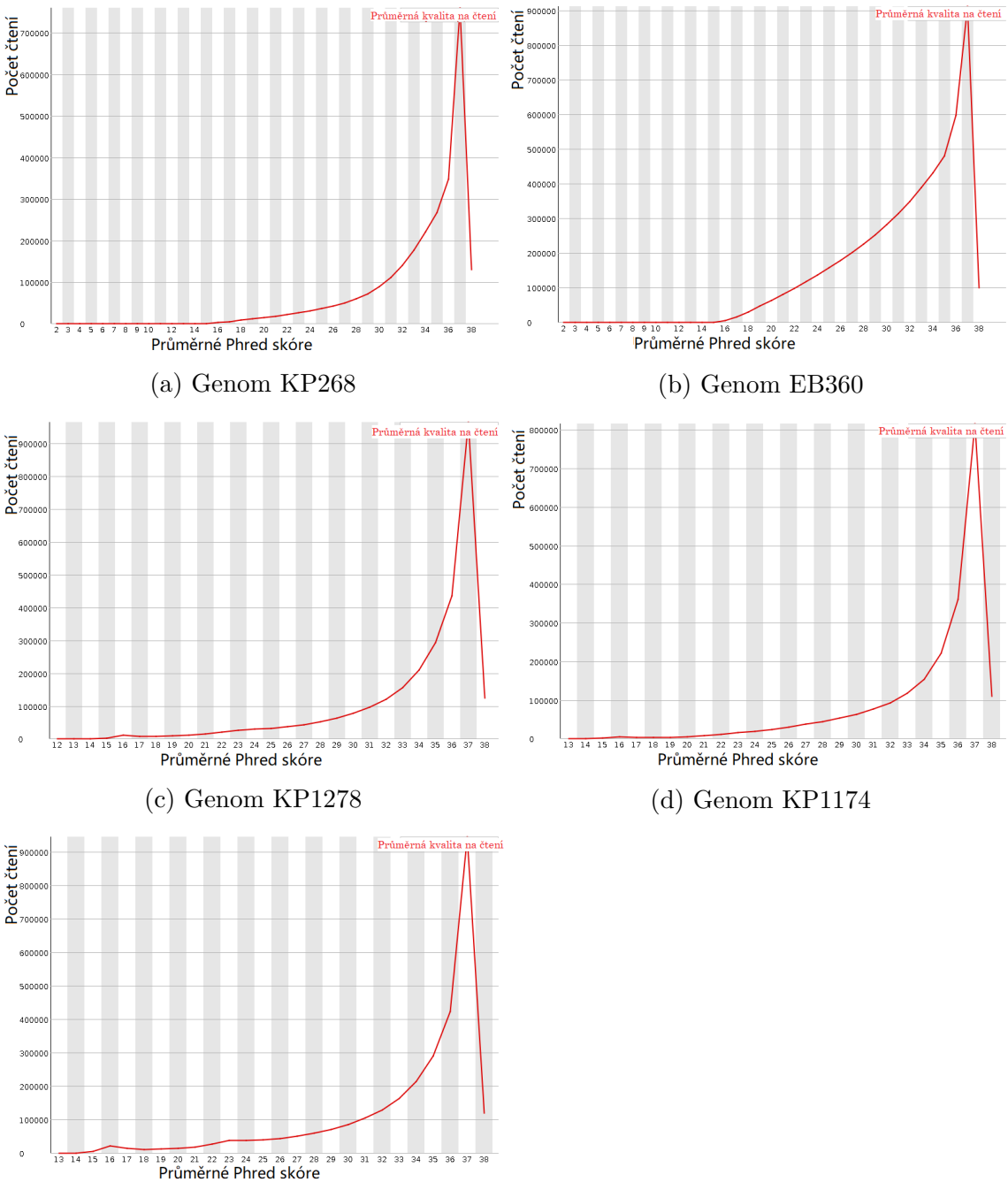


(d) Genom KP1174



(e) Genom KP1268

Obr. A.1: Seskupení skóre kvality v každé pozici jednotlivých čtení genomů sekvenovaných pomocí Illumina Miseq.



(a) Genom KP268

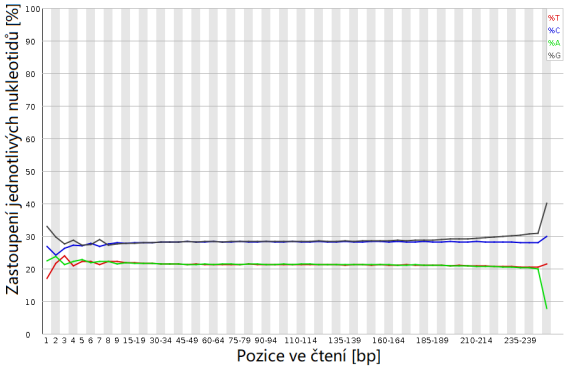
(b) Genom EB360

(c) Genom KP1278

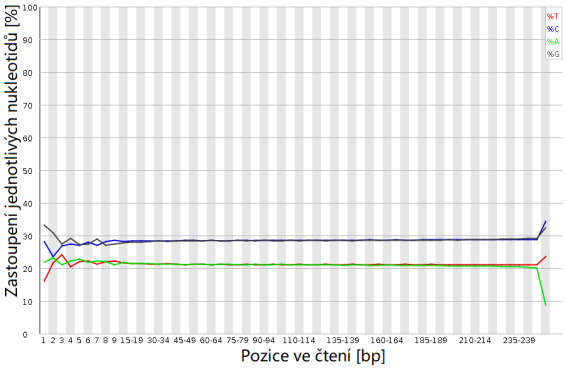
(d) Genom KP1174

(e) Genom KP1268

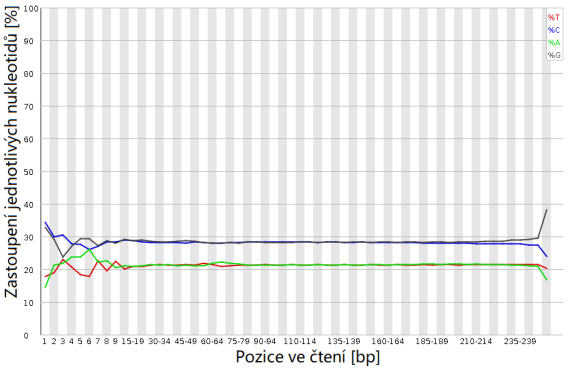
Obr. A.2: Distribuce Phred skóre na celkový počet čtení sekvence genomů sekvenovaných pomocí Illumina Miseq.



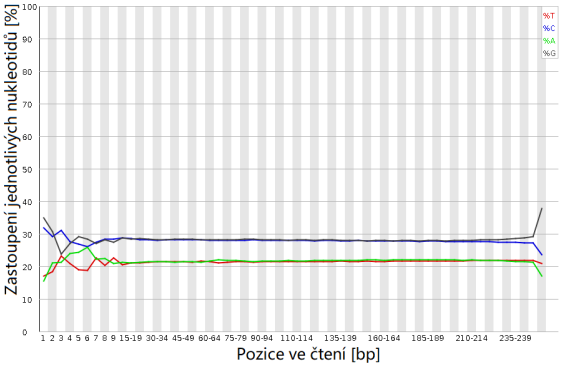
(a) Genom KP268



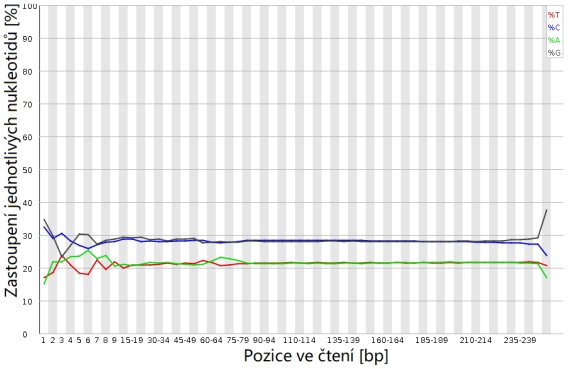
(b) Genom EB360



(c) Genom KP1278

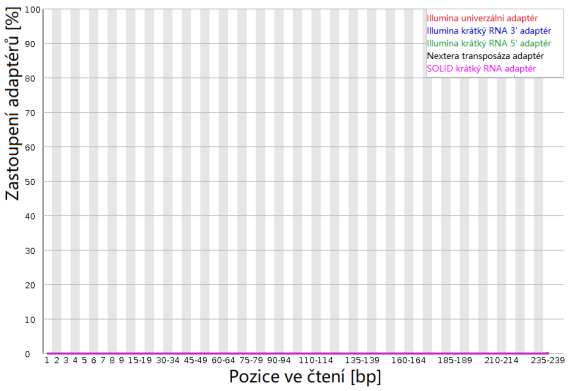


(d) Genom KP1174

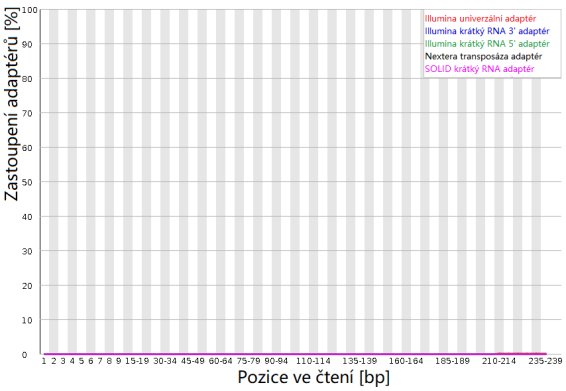


(e) Genom KP1268

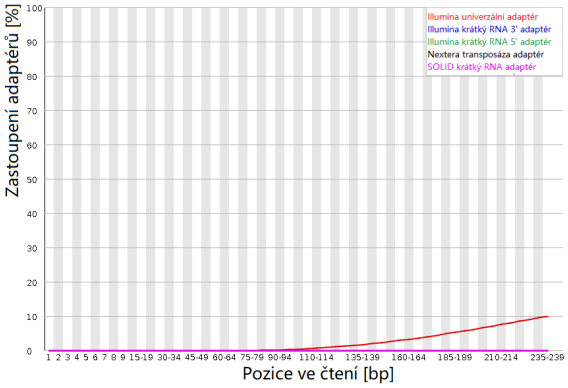
Obr. A.3: Procentuální zastoupení přečtených bází pro každý ze čtyř nukleotidů v každé pozici jednotlivých čtení genomů sekvenovaných pomocí Illumina Miseq.



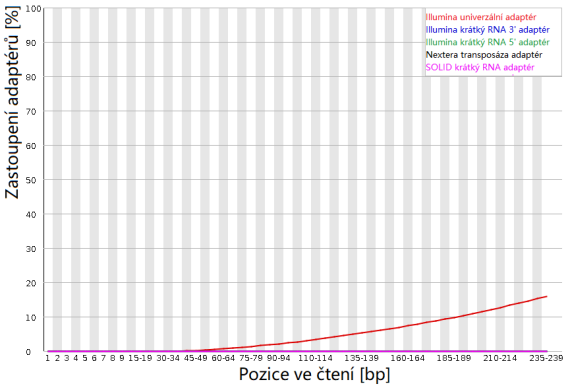
(a) Genom KP268



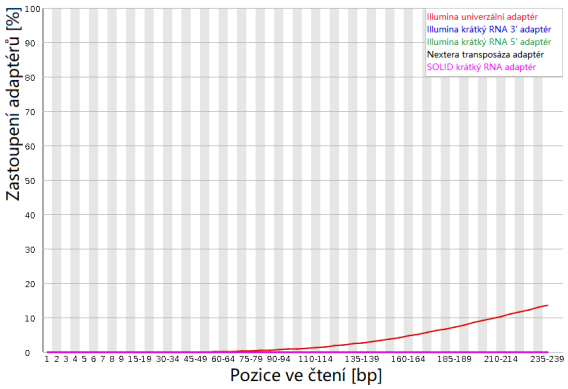
(b) Genom EB360



(c) Genom KP1278

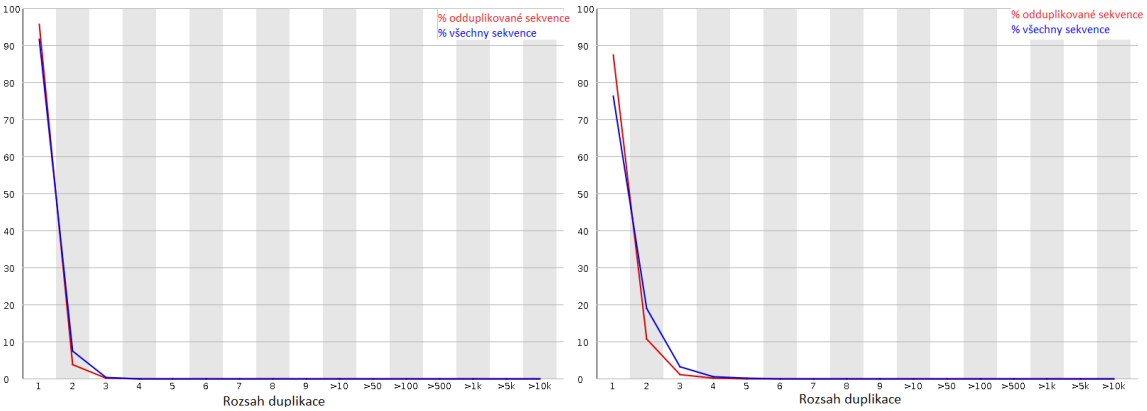


(d) Genom KP1174



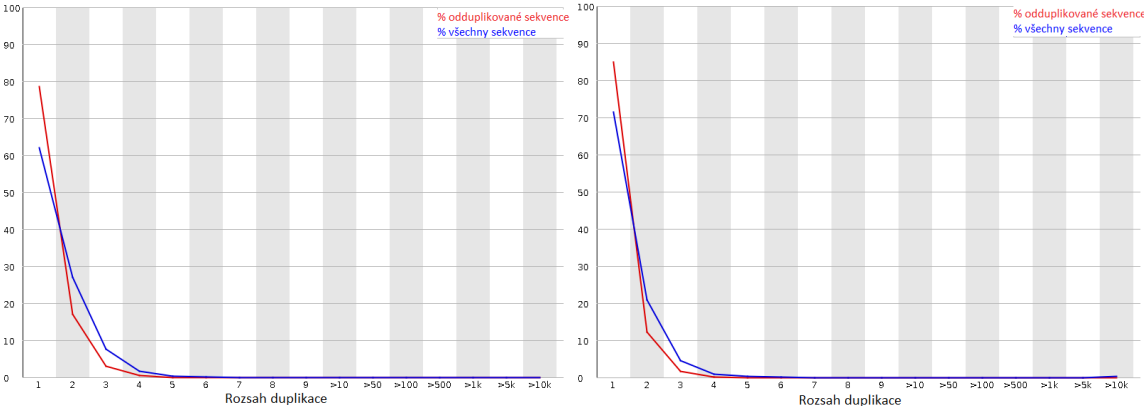
(e) Genom KP1268

Obr. A.4: Procentuální zastoupení adaptérů v každé pozici jednotlivých čtení genomů sekvenovaných pomocí Illumina Miseq.



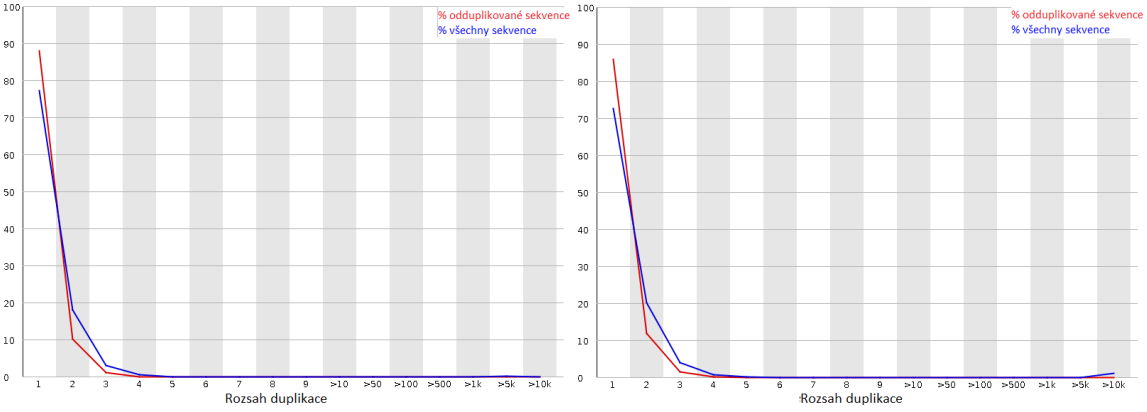
(a) Genom EB359

(b) Genom KP268



(c) Genom EB360

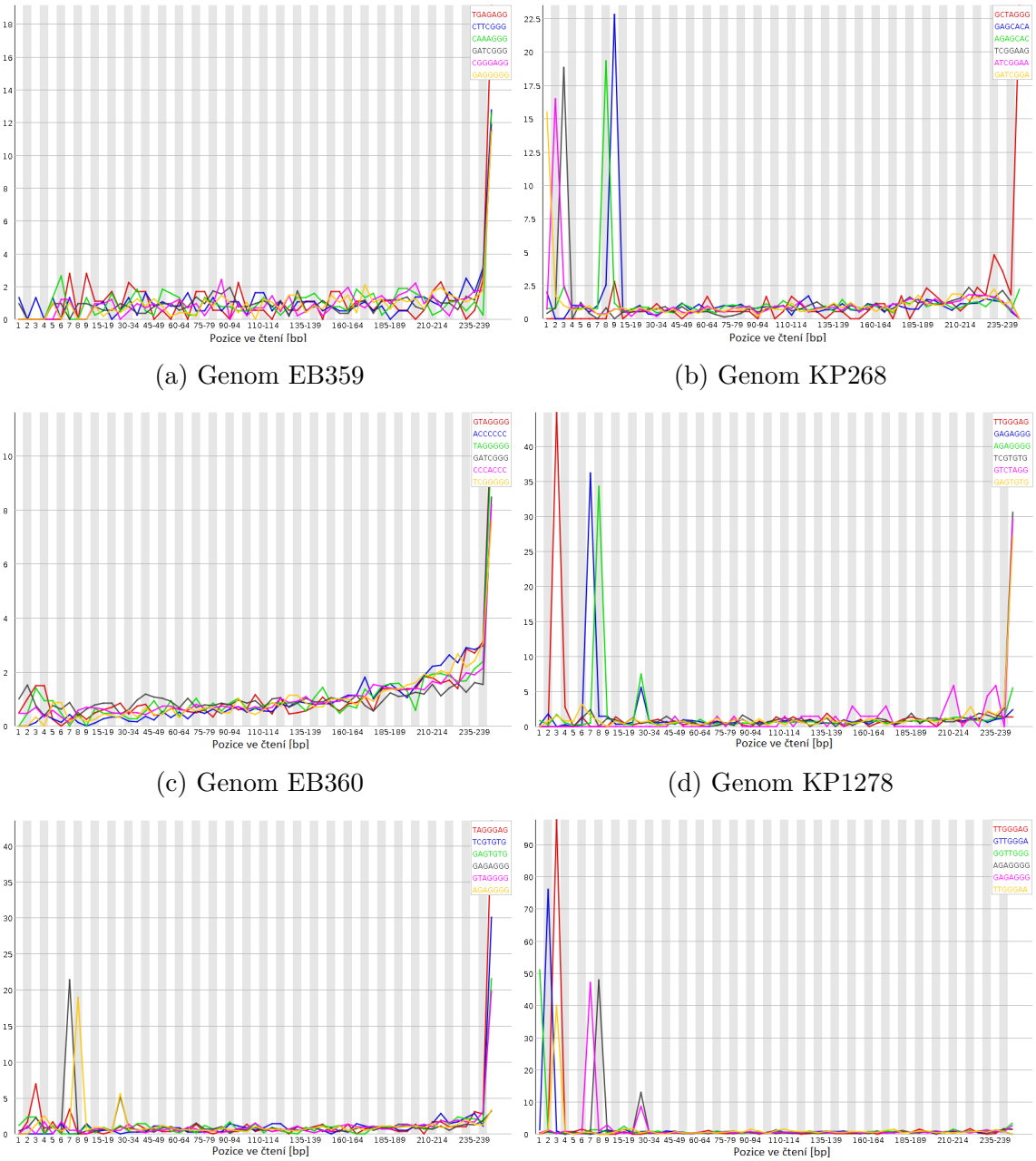
(d) Genom KP1278



(e) Genom KP1174

(f) Genom KP1268

Obr. A.5: Stupeň duplikace v genomech sekvenovaných pomocí Illumina Miseq.



(a) Genom EB359

(b) Genom KP268

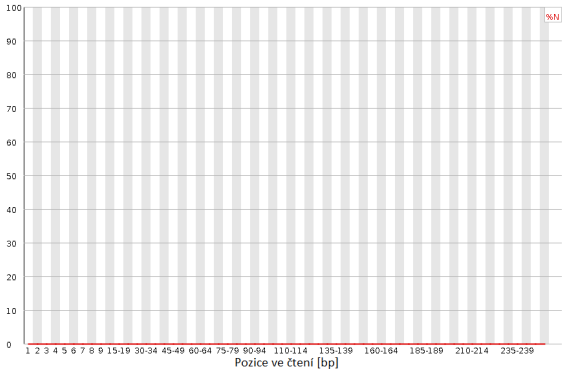
(c) Genom EB360

(d) Genom KP1278

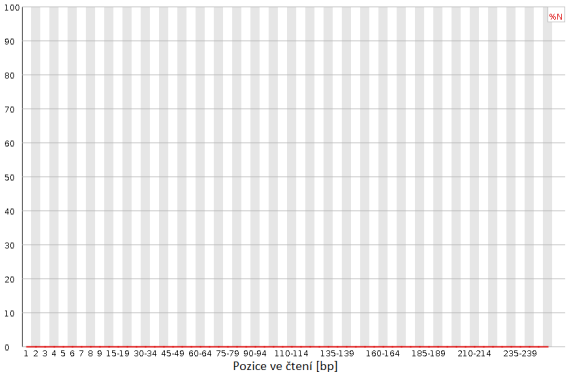
(e) Genom KP1174

(f) Genom KP1268

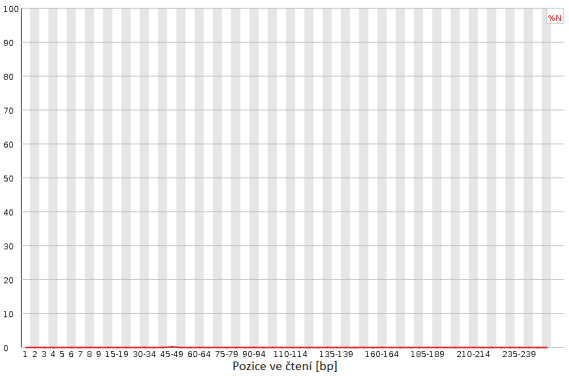
Obr. A.6: K-mery v genomech sekvenovaných pomocí Illumina Miseq.



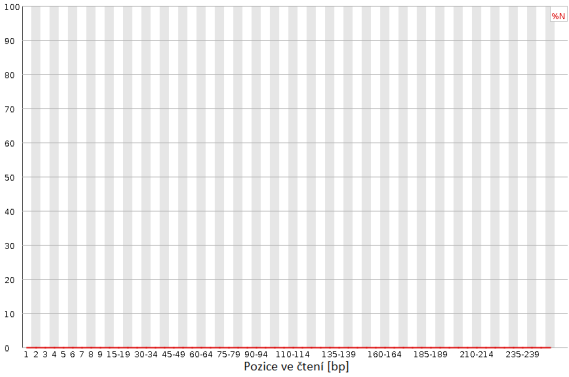
(a) Genom EB359



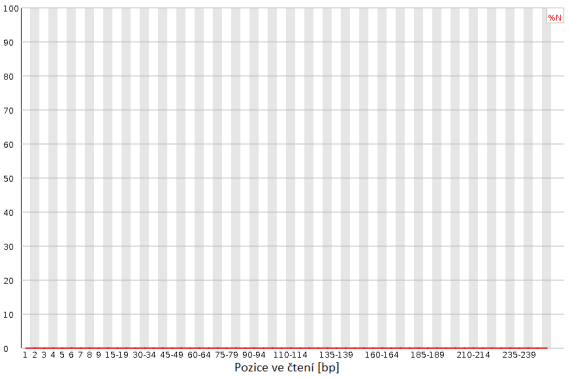
(b) Genom KP268



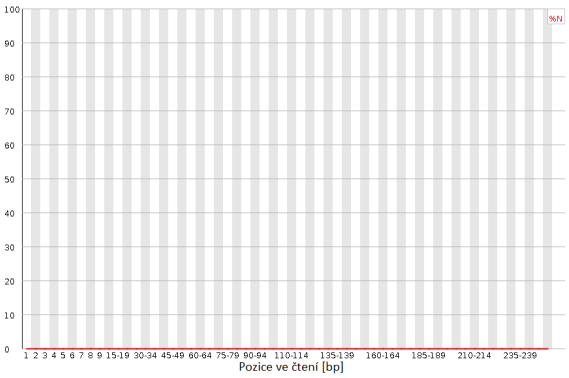
(c) Genom EB360



(d) Genom KP1278

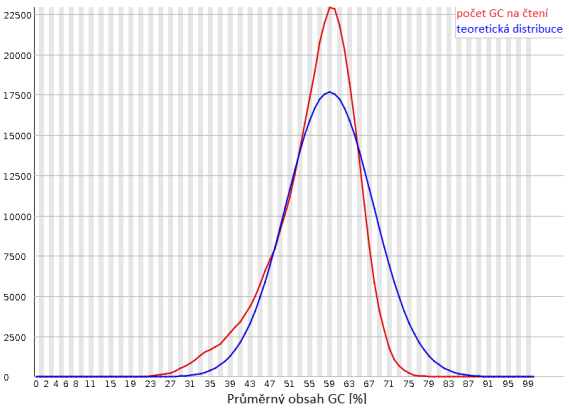


(e) Genom KP1174

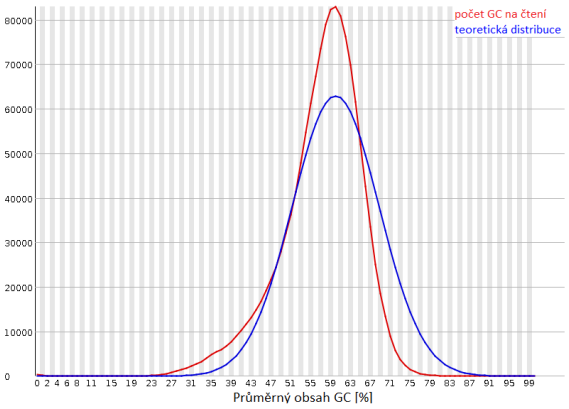


(f) Genom KP1268

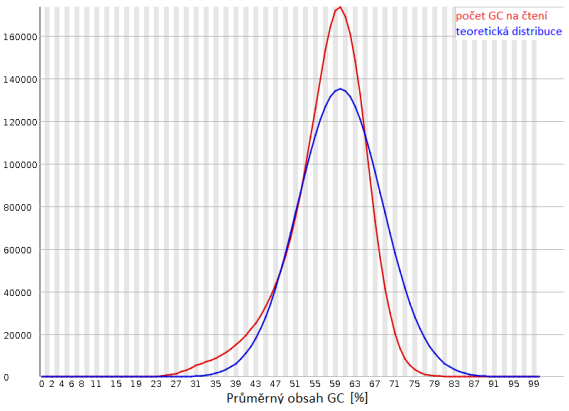
Obr. A.7: Obsah N v genomech sekvenovaných pomocí Illumina Miseq.



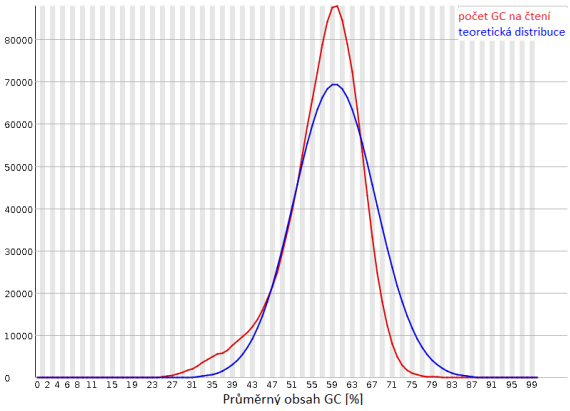
(a) Genom EB359



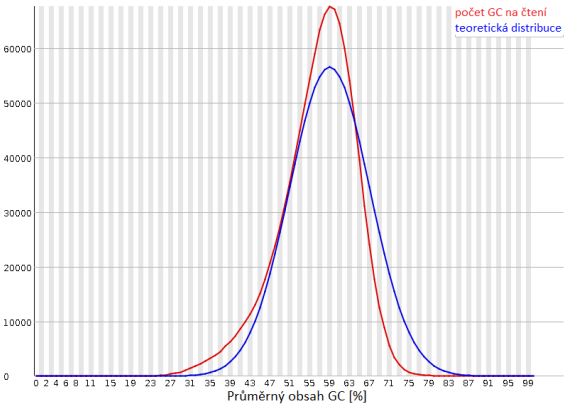
(b) Genom KP268



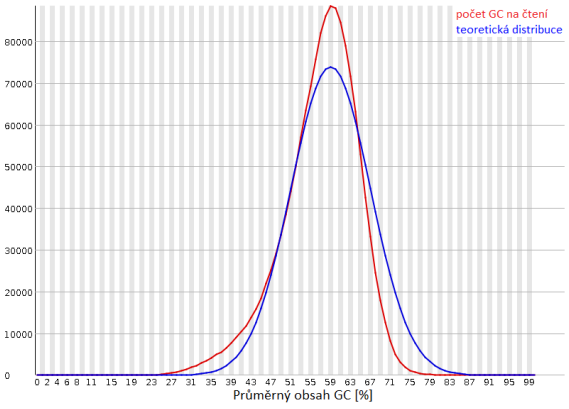
(c) Genom EB360



(d) Genom KP1278

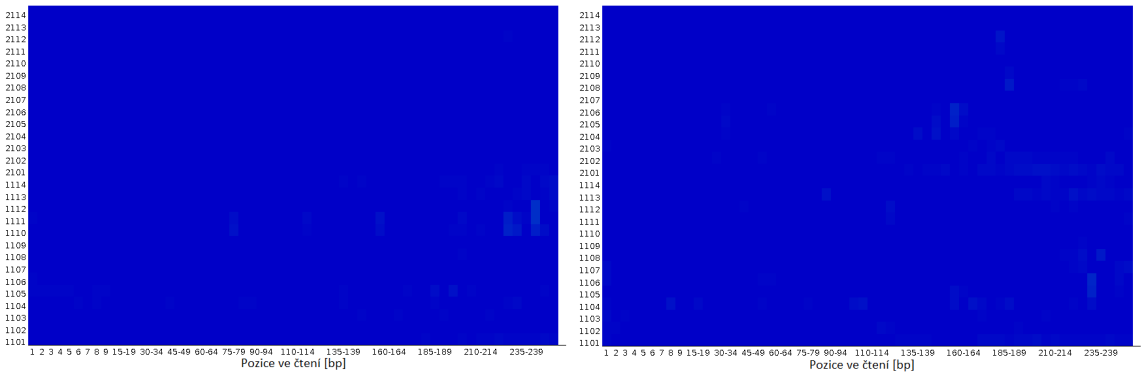


(e) Genom KP1174



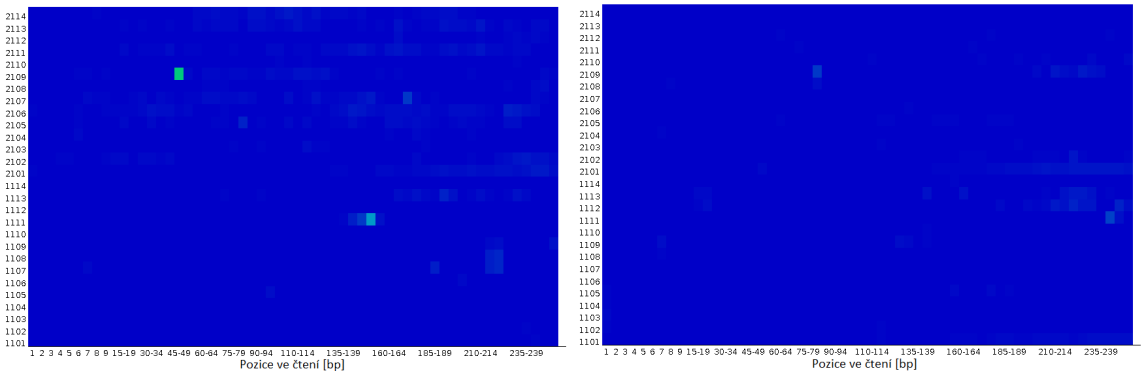
(f) Genom KP1268

Obr. A.8: Průměrný obsah bází G a C v genomech sekvenovaných pomocí Illumina Miseq.



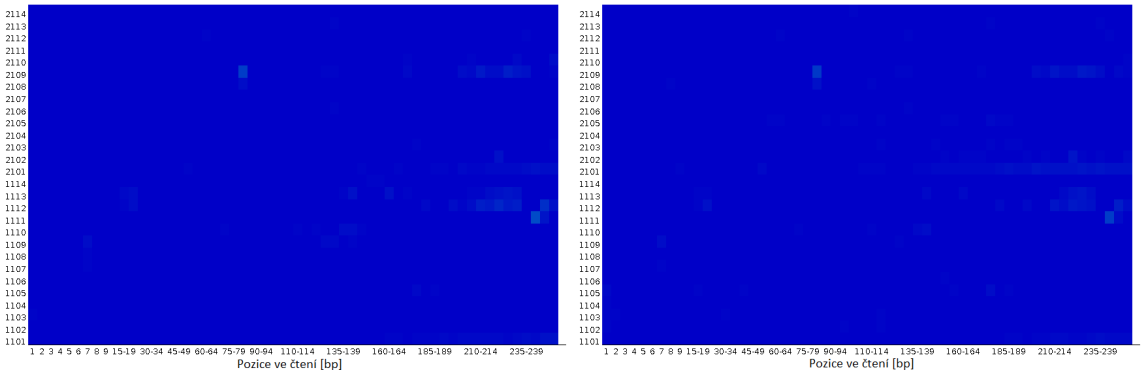
(a) Genom EB359

(b) Genom KP268



(c) Genom EB360

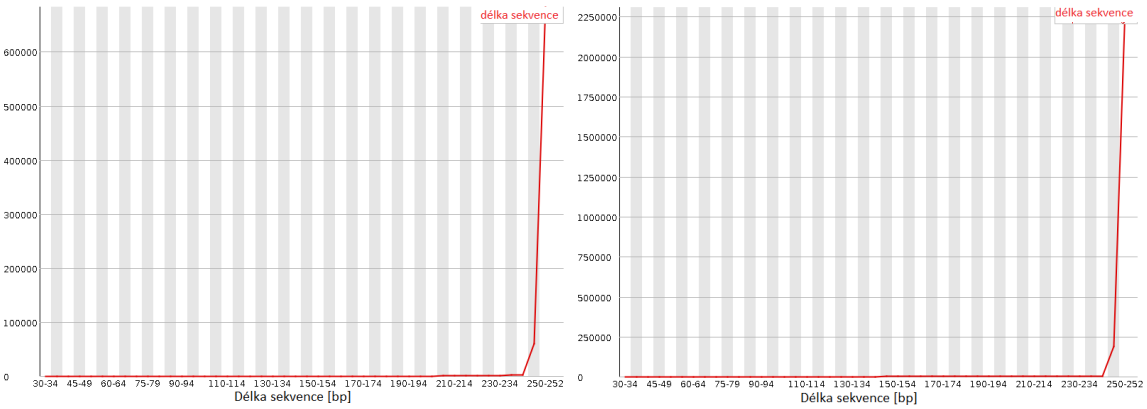
(d) Genom KP1278



(e) Genom KP1174

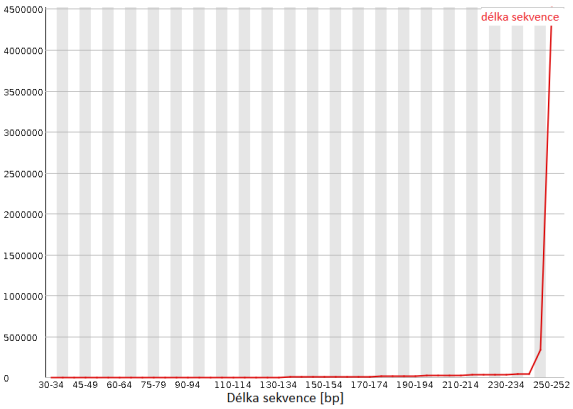
(f) Genom KP1268

Obr. A.9: Kvalita sekvenace na pozici v rámci destičky sekvenátoru Illumina Miseq.

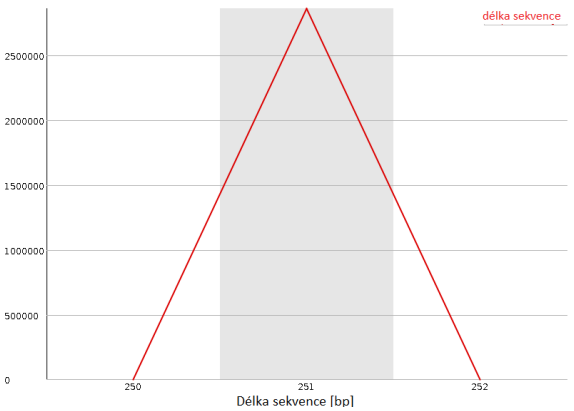


(a) Genom EB359

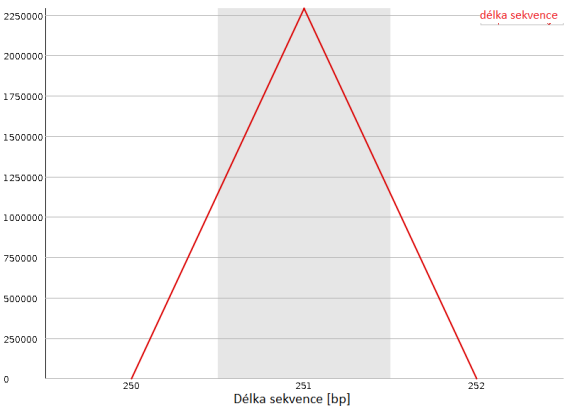
(b) Genom KP268



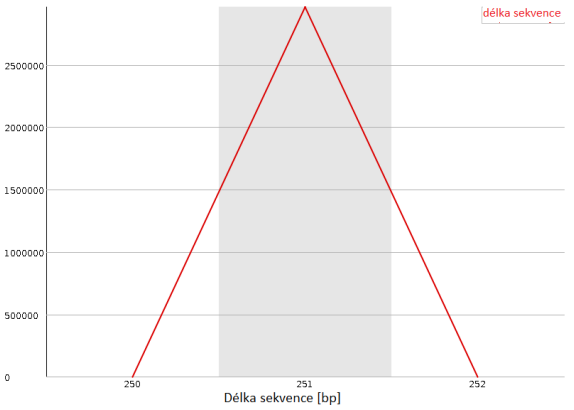
(c) Genom EB360



(d) Genom KP1278



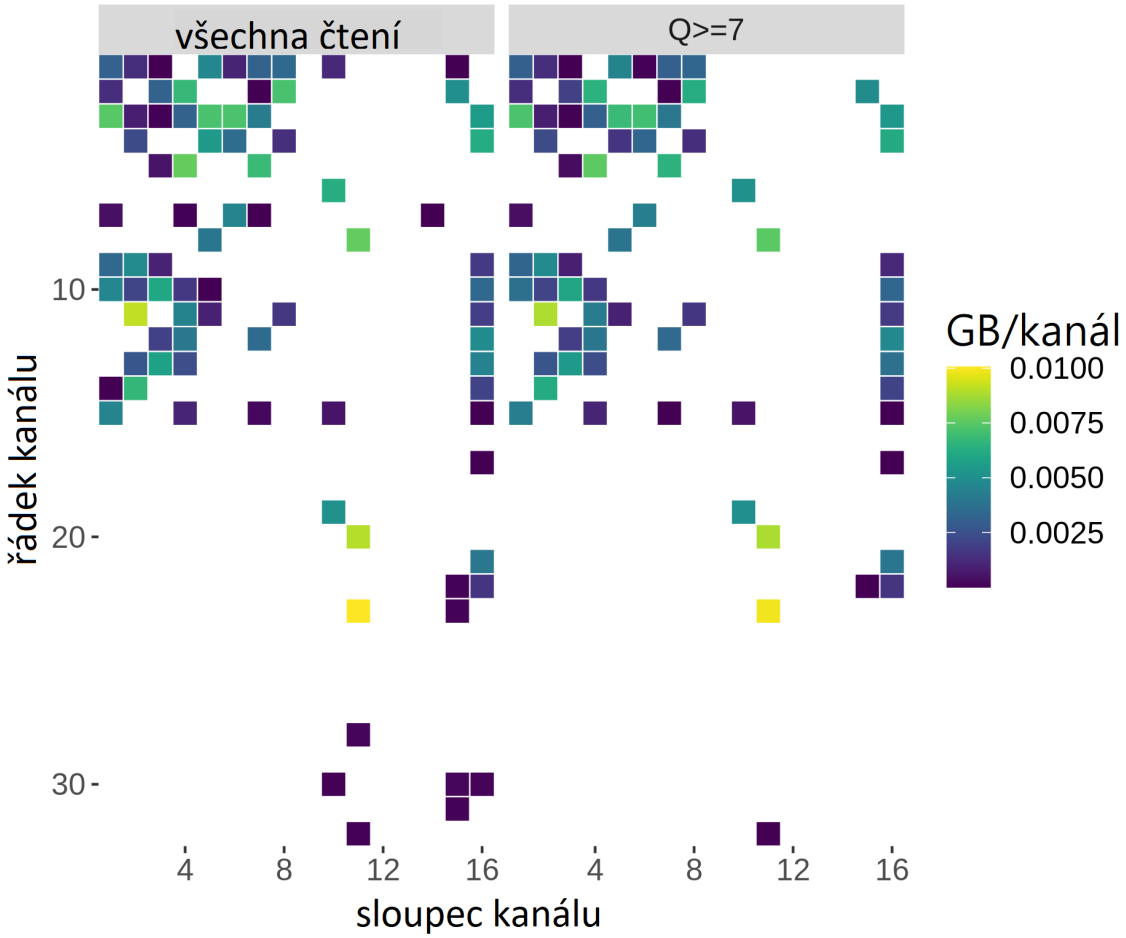
(e) Genom KP1174



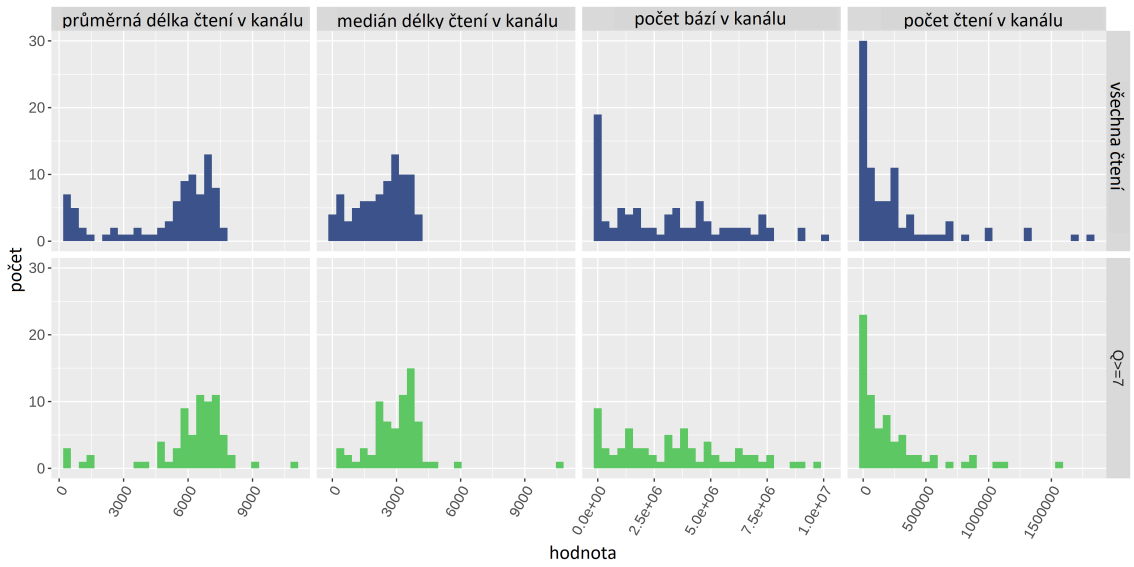
(f) Genom KP1268

Obr. A.10: Distribuce délky sekvence v genomech sekvenovaných pomocí Illumina Miseq.

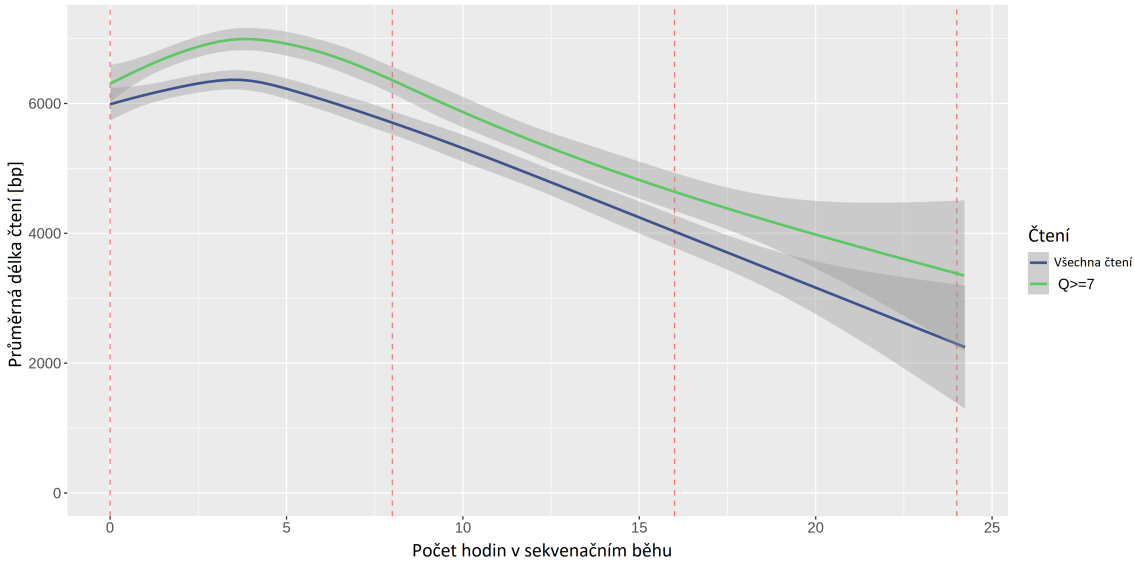
B Výstup z MinIONQC



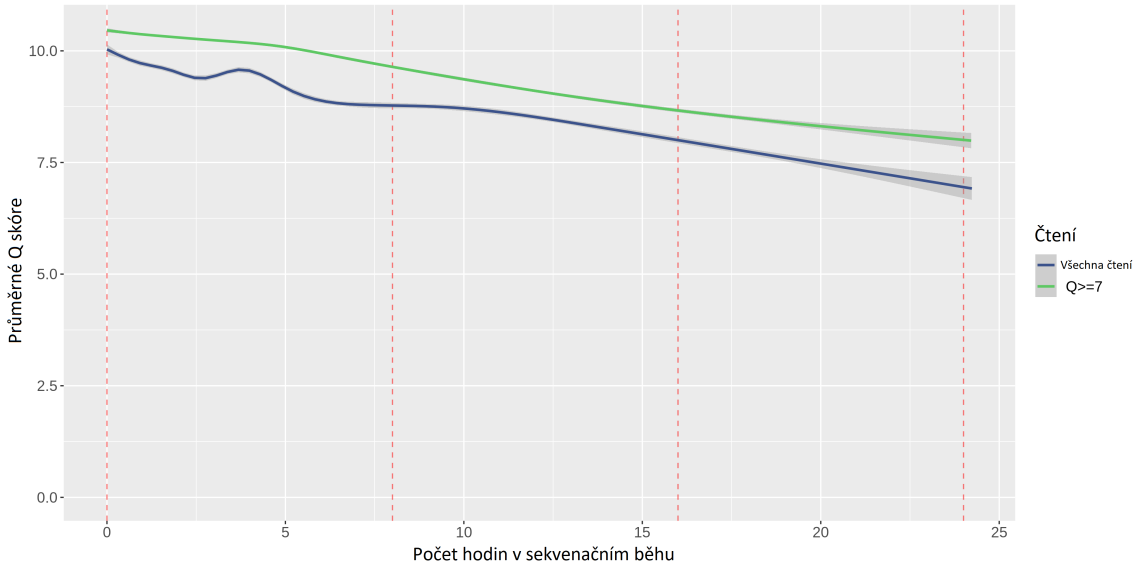
Obr. B.1: Počet gigabází sekvenovaných v každém kanálu sekvenační komůrky platformy MinION.



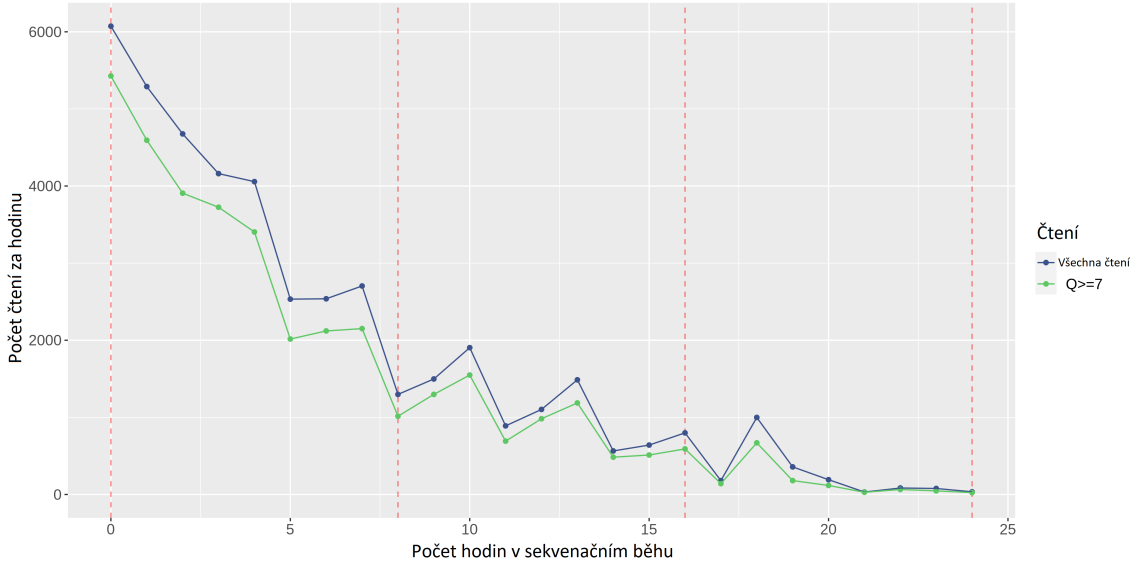
Obr. B.2: Histogramy celkových bází, čtení, průměru a mediánu délek čtení sekvenovaných na platformě MinION.



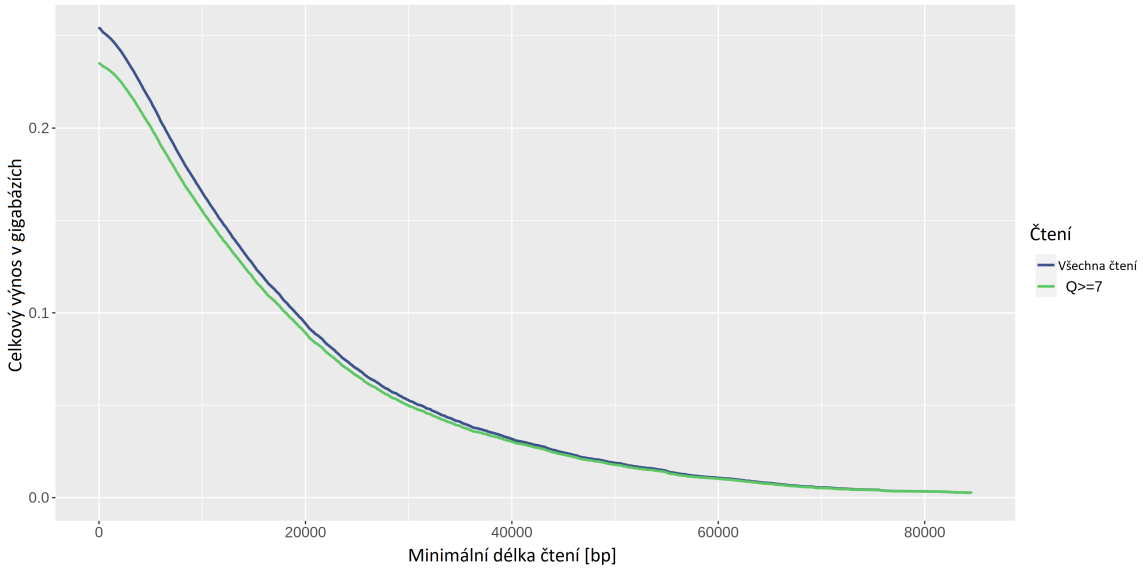
Obr. B.3: Průměrné délky čtení sekvenovaných na platformě MinION v průběhu jednoho běhu.



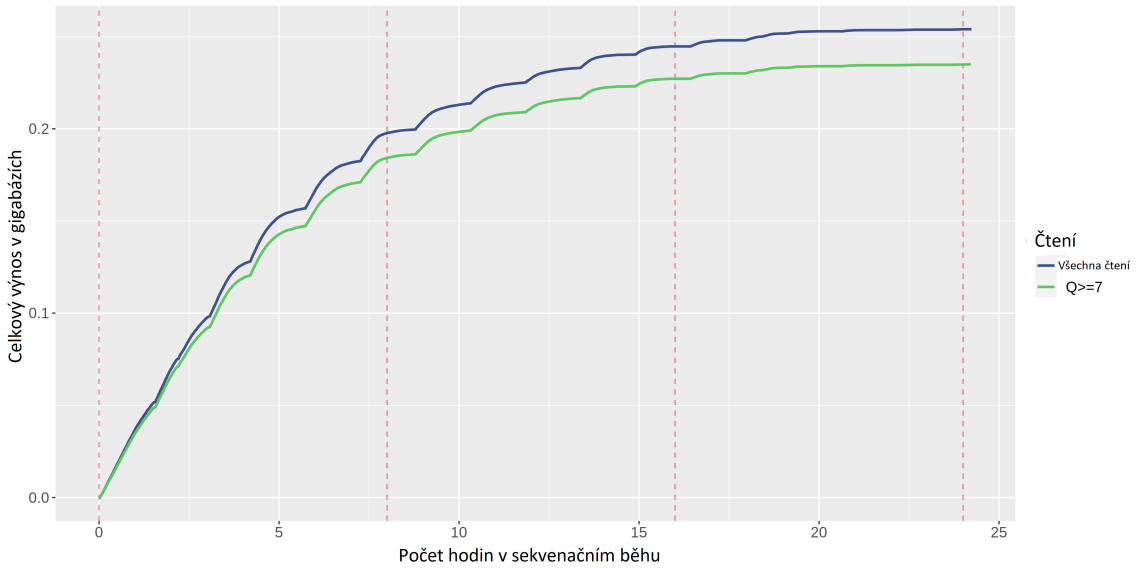
Obr. B.4: Průměrné skóre kvality čtení sekvenovaných na platformě MinION v průběhu jednoho běhu.



Obr. B.5: Počet čtení produkovaných na platformě MinION v průběhu jednoho běhu.

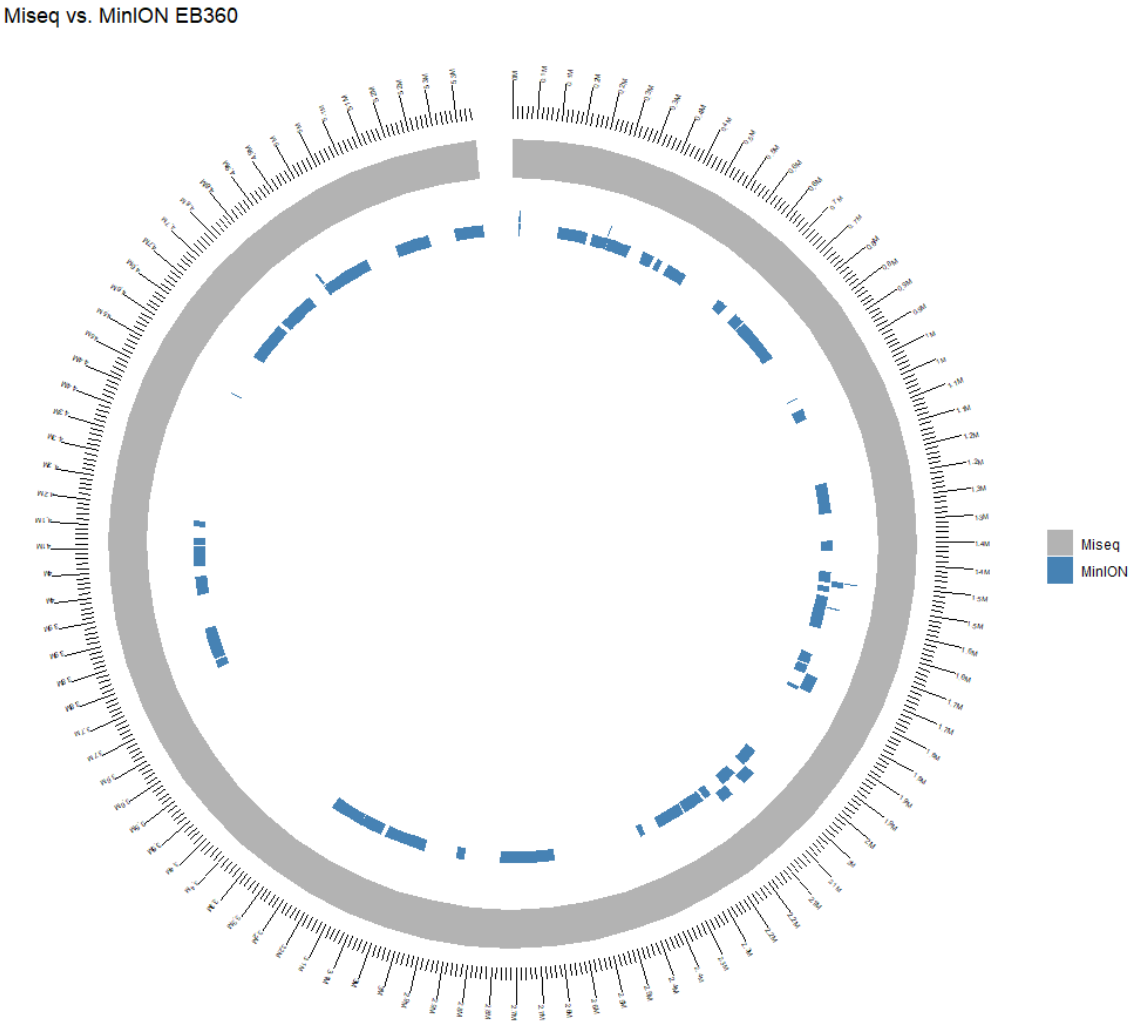


Obr. B.6: Celková produkce gigabází pro minimální délku čtení sekvenovaných na platformě MinION.

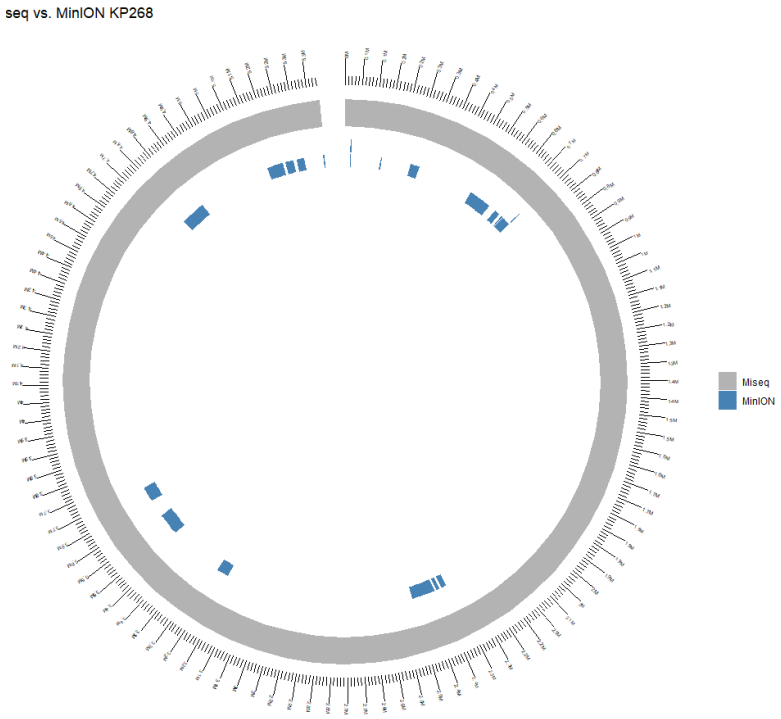


Obr. B.7: Celková produkce gigabází na platformě MinION v průběhu jednoho běhu.

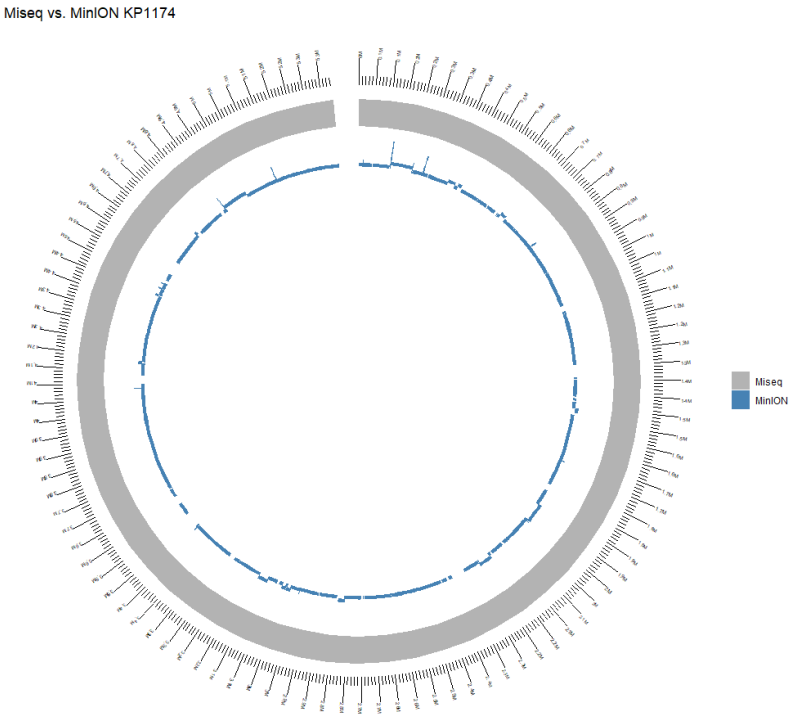
C Grafy zarovnání sestavených sekvencí



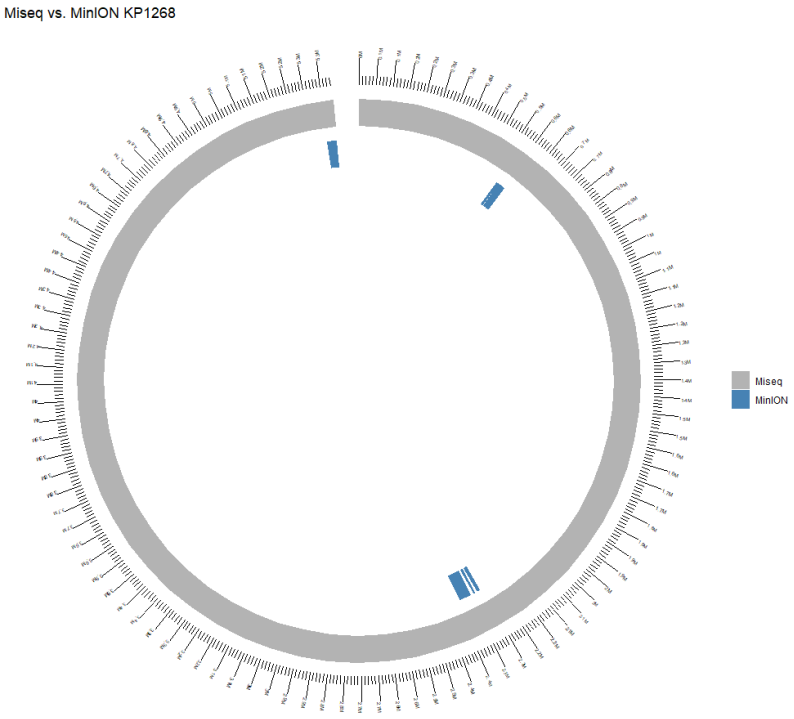
Obr. C.1: Grafické znázornění zarovnání sekvence EB360 sekvenované pomocí platformy Miseq a sekvence EB360 sekvenované na platformě MinION.



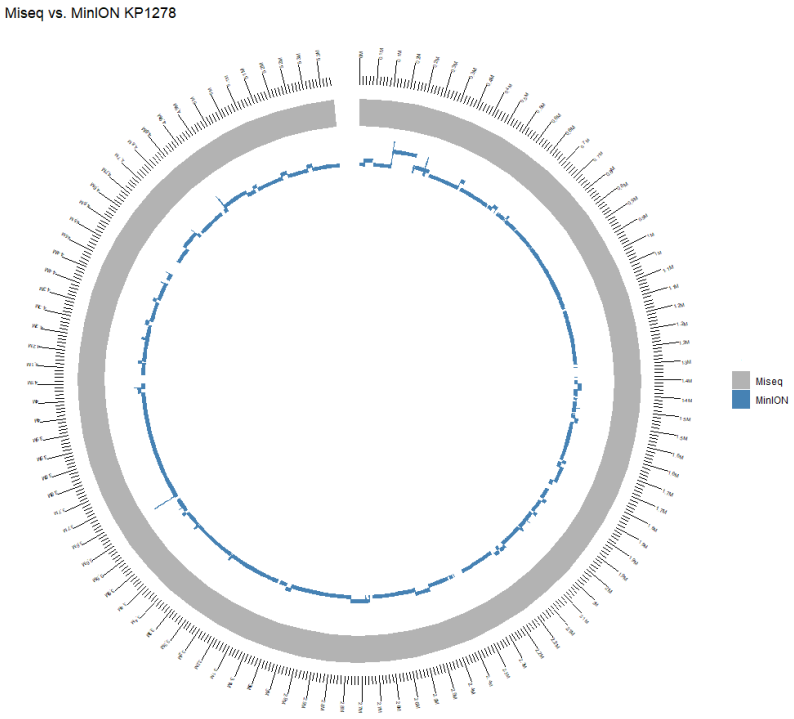
Obr. C.2: Grafické znázornění zarovnání sekvence KP268 sekvenované pomocí platformy Miseq a sekvence KP268 sekvenované na platformě MinION.



Obr. C.3: Grafické znázornění zarovnání sekvence KP1174 sekvenované pomocí platformy Miseq a sekvence KP1174 sekvenované na platformě MinION.

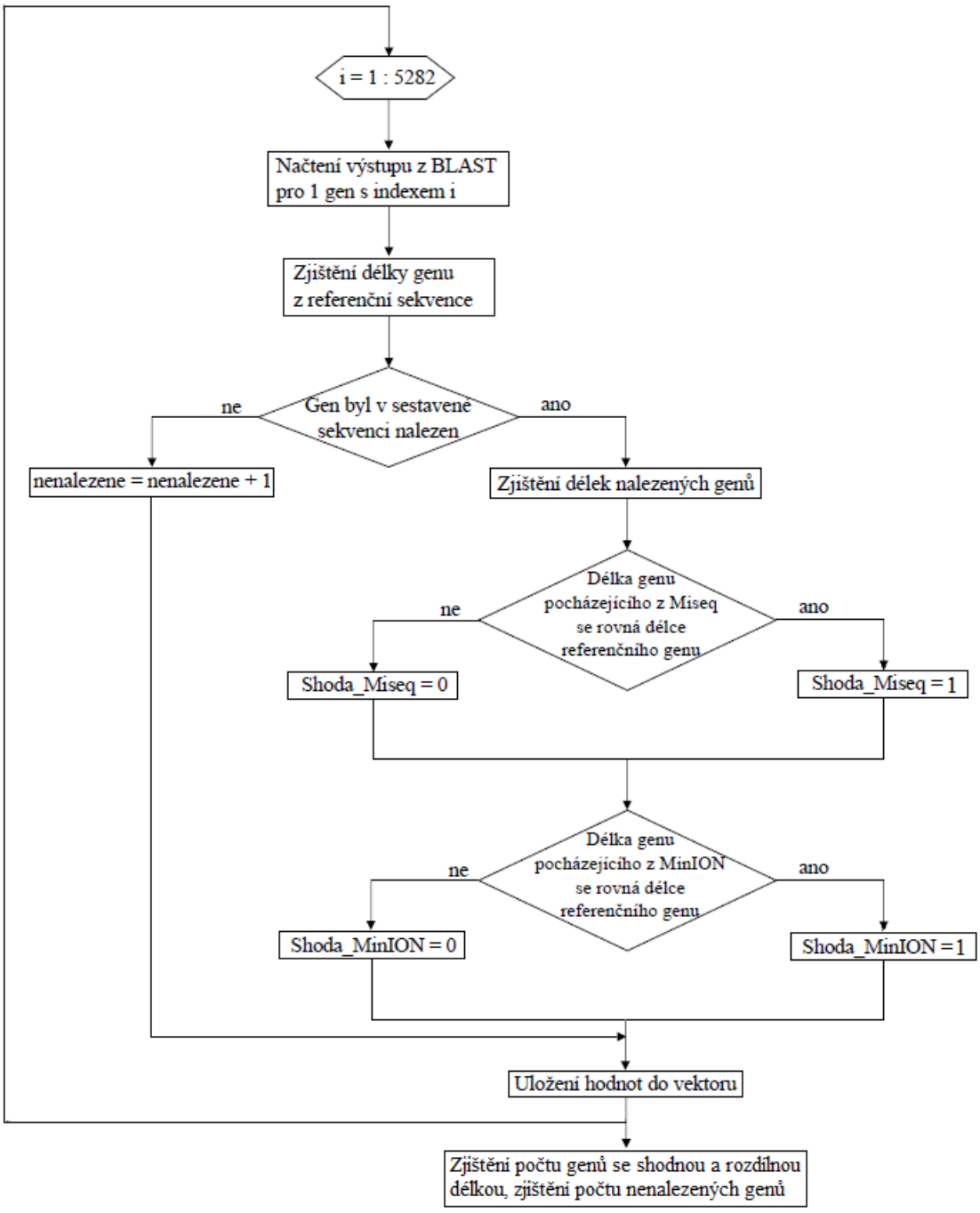


Obr. C.4: Grafické znázornění zarovnání sekvence KP1268 sekvenované pomocí platformy Miseq a sekvence KP1268 sekvenované na platformě MinION.

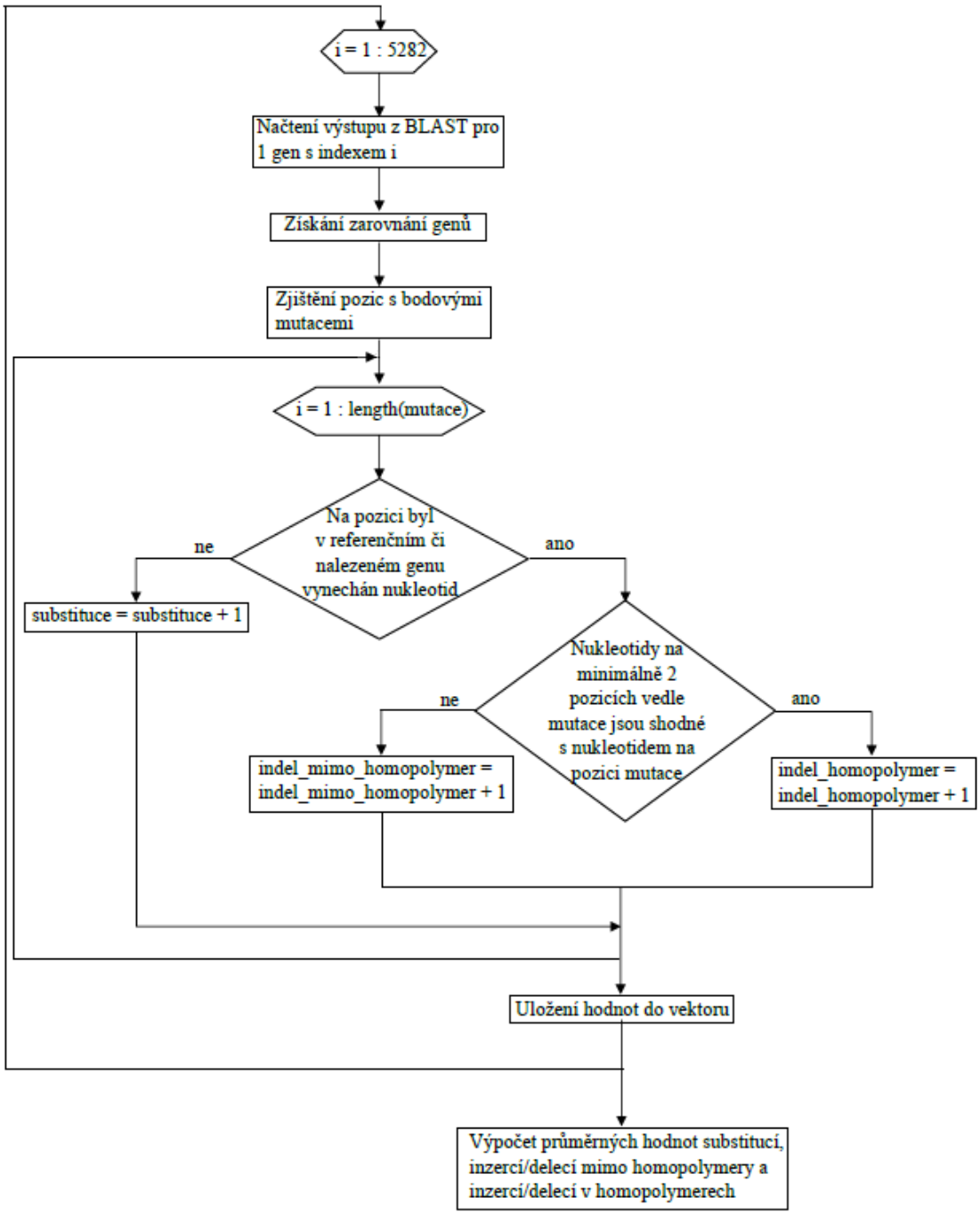


Obr. C.5: Grafické znázornění zarovnání sekvence KP1278 sekvenované pomocí platformy Miseq a sekvence KP1278 sekvenované na platformě MinION.

D Bloková schémata vytvořených algoritmů

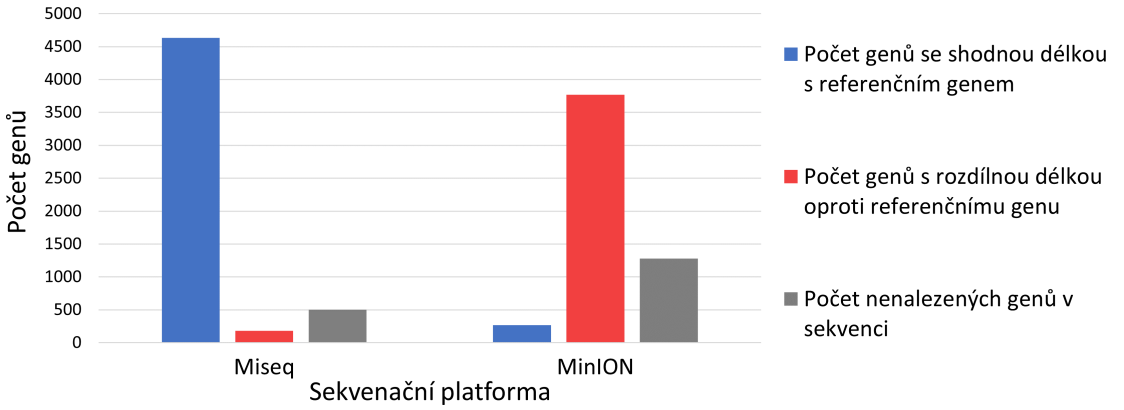


Obr. D.1: Blokové schéma algoritmu pro porovnání délek nalezených genů.

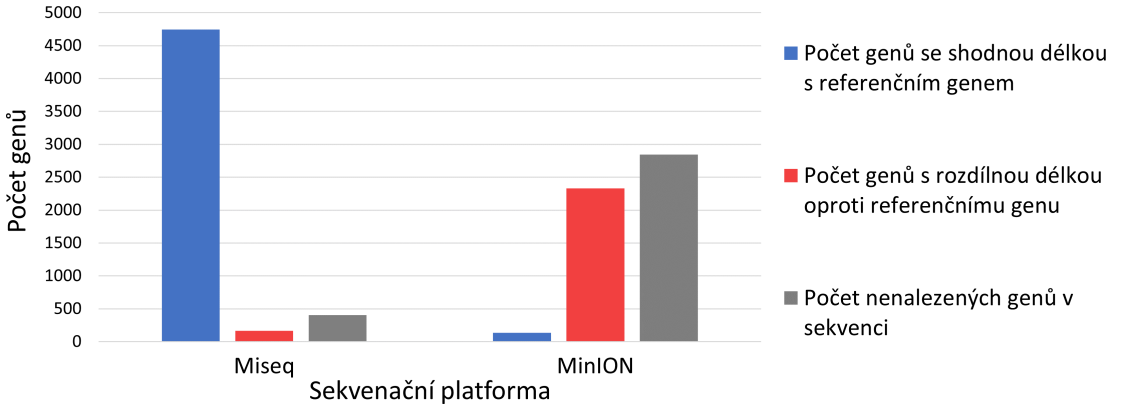


Obr. D.2: Blokové schéma algoritmu pro analýzu bodových mutací v nalezených genech.

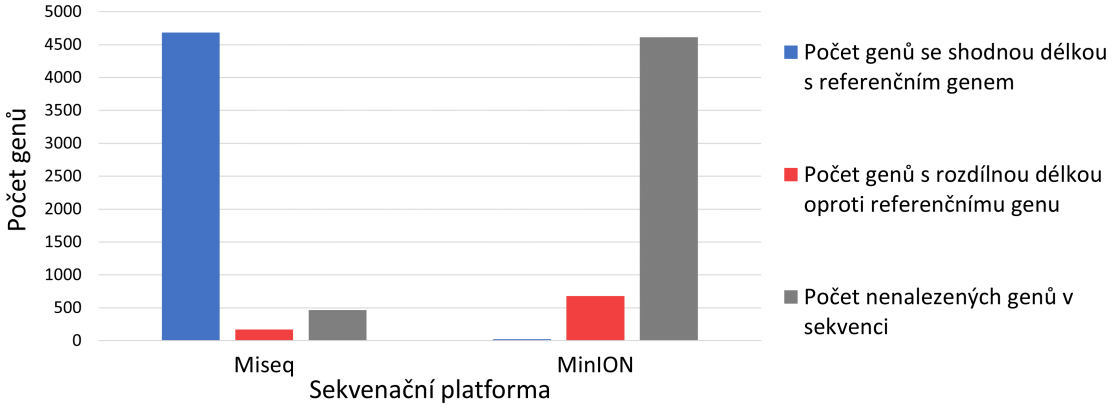
E Grafy porovnávání genomů na základě nalezených genů



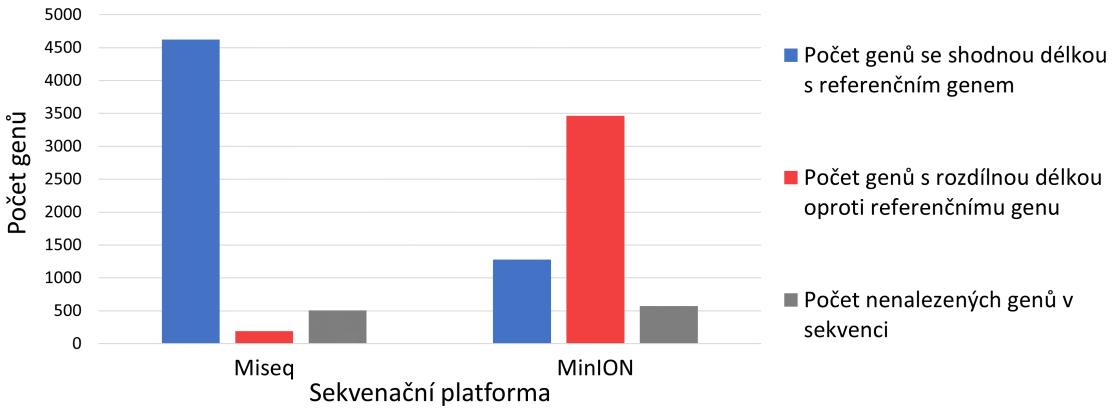
Obr. E.1: Porovnání počtu genů nalezených v genomu EB359 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



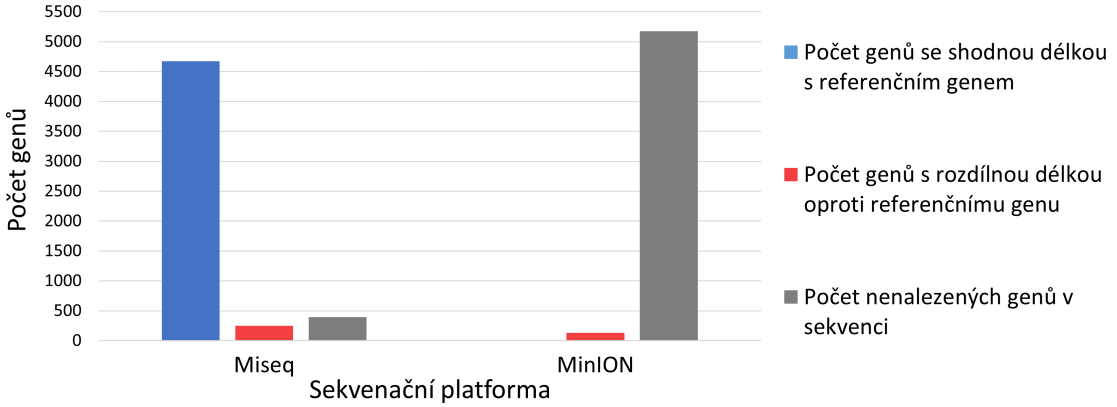
Obr. E.2: Porovnání počtu genů nalezených v genomu EB360 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



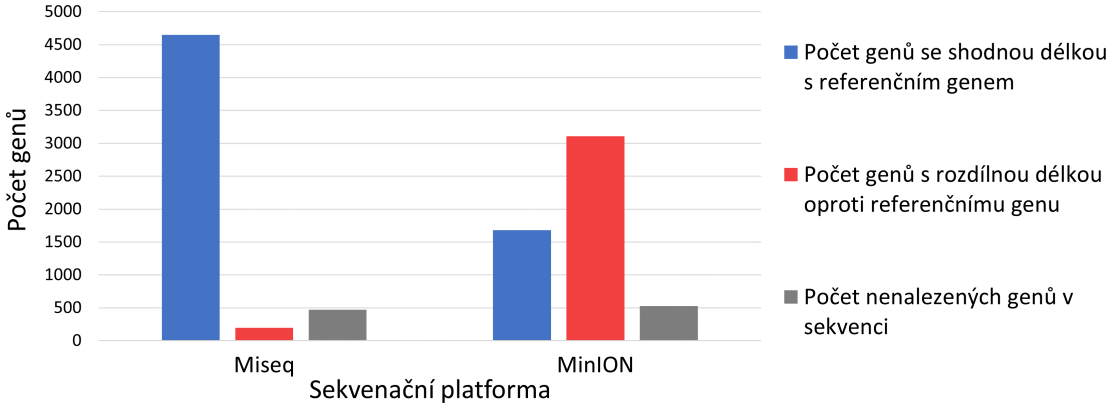
Obr. E.3: Porovnání počtu genů nalezených v genomu KP268 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



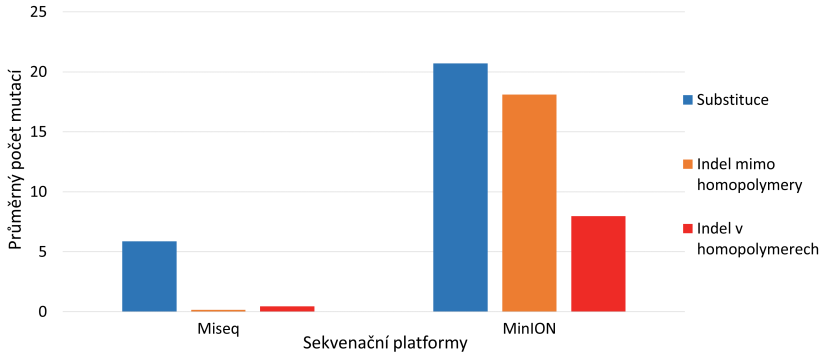
Obr. E.4: Porovnání počtu genů nalezených v genomu KP1174 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



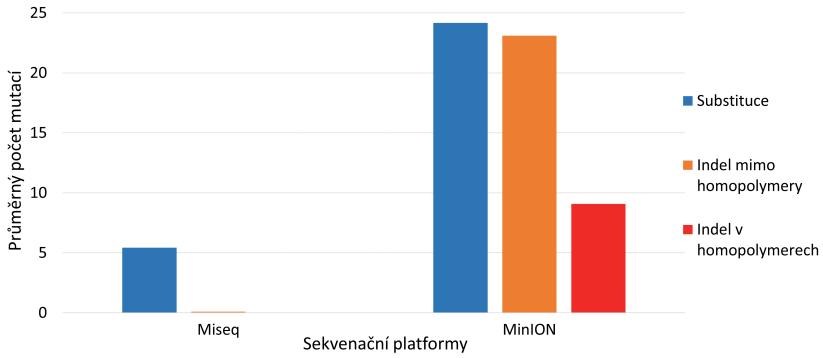
Obr. E.5: Porovnání počtu genů nalezených v genomu KP1268 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



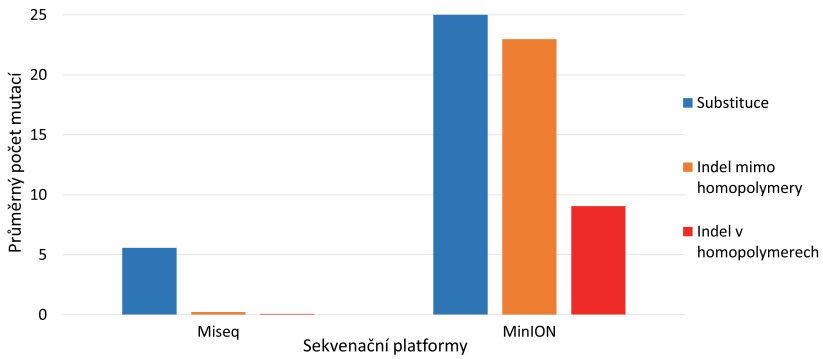
Obr. E.6: Porovnání počtu genů nalezených v genomu KP1278 sekvenovaných na platformě Miseq a MinION z hlediska jejich délek.



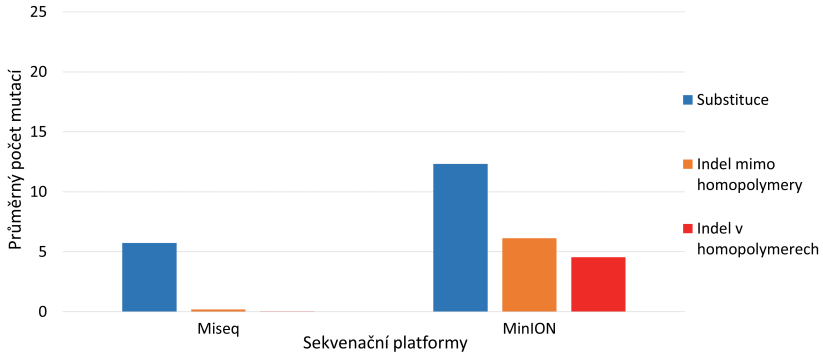
Obr. E.7: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí EB359 obdržených z dvou druhů sekvenátorů.



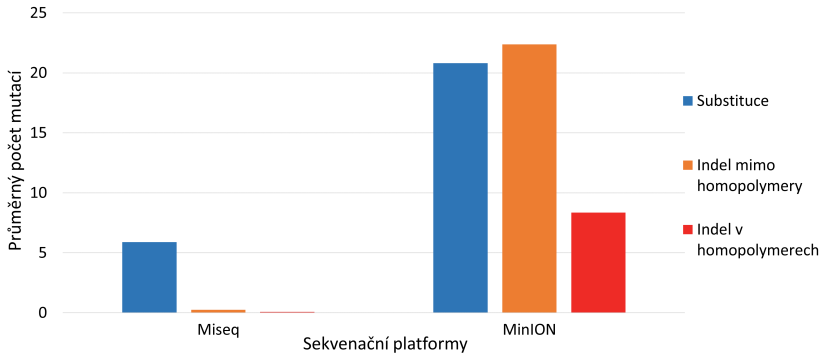
Obr. E.8: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí EB360 obdržených z dvou druhů sekvenátorů.



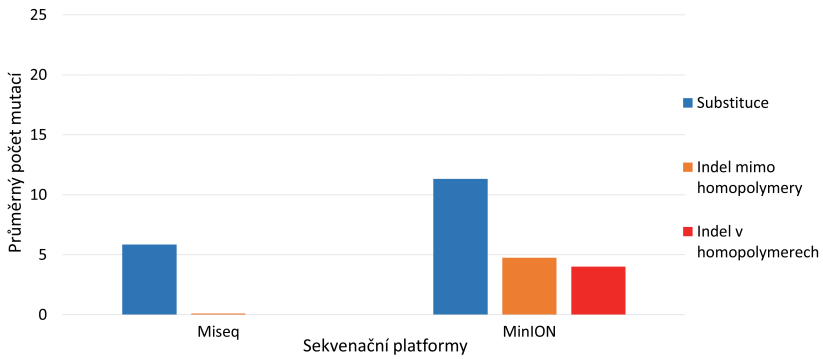
Obr. E.9: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí KP268 obdržených z různých druhů sekvenátorů.



Obr. E.10: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí KP1174 obdržených z různých druhů sekvenátorů.



Obr. E.11: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí KP1268 obdržených z různých druhů sekvenátorů.



Obr. E.12: Graf porovnání průměrného počtu bodových mutací v nalezených genech sekvencí KP1278 obdržených z různých druhů sekvenátorů.

F Tabulky hodnot porovnávání genomů na základě nalezených genů

Tab. F.1: Hodnoty porovnání délek nalezených genů s délkami genů obsažených v referenční sekvenci.

ID genomu	Sekvenátor	Počet genů se shodnou délkou s referenční sekvencí	Počet genů s rozdílnou délkou oproti referenčnímu genu	Počet nenalezených genů v sekvenci
EB359	Miseq	4 600	183	499
	MinION	266	3 737	1 279
EB360	Miseq	4 709	167	406
	MinION	138	2 299	2 845
KP268	Miseq	4 648	170	464
	MinION	25	644	4 613
KP1174	Miseq	4 588	189	505
	MinION	1 280	3 430	572
KP1268	Miseq	4 640	248	394
	MinION	8	97	5 177
KP1278	Miseq	4 617	196	469
	MinION	1 680	3 074	528

Tab. F.2: Hodnoty porovnání nalezených genů z hlediska bodových mutací.

ID genomu	Sekvenátor	Substituce	Indel mimo homopolymery	Indel v homopolymerech
EB359	Miseq	5,88	0,17	0,46
	MinION	20,71	18,11	7,97
EB360	Miseq	5,42	0,1	0,03
	MinION	24,17	23,10	9,07
KP268	Miseq	5,58	0,23	0,07
	MinION	25,11	22,98	9,06
KP1174	Miseq	5,74	0,19	0,05
	MinION	12,32	6,13	4,56
KP1268	Miseq	5,9	0,25	0,07
	MinION	20,81	22,37	8,35
KP1278	Miseq	5,86	0,11	0,03
	MinION	11,32	4,76	4,01